ESCAPE: Encoding Super-keypoints for Category-Agnostic Pose Estimation

Khoi Duc Nguyen[1]* Chen Li[2] Gim Hee Lee[2]
[1]University of Wisconsin-Madison [2]National University of Singapore
khoi.nguyen@wisc.edu lichen@u.nus.edu gimhee.lee@comp.nus.edu.sg

Abstract

In this paper, we tackle the task of category-agnostic pose estimation (CAPE), which aims to predict poses for objects of any category with few annotated samples. Previous works either rely on local matching between features of support and query samples or require support keypoint identifier. The former is prone to overfitting due to its sensitivity to sparse samples, while the latter is impractical for the open-world nature of the task. To overcome these limitations, we propose ESCAPE - a Bayesian framework that learns a prior over the features of keypoints. The prior can be expressed as a mixture of super-keypoints, each being a high-level abstract keypoint that captures the statistics of semantically related keypoints from different categories. We estimate the super-keypoints from base categories and use them in adaptation to novel categories. The adaptation to an unseen category involves two steps: first, we match each novel keypoint to a related super-keypoint; and second, we transfer the knowledge encoded in the matched super-keypoints to the novel keypoints. For the first step, we propose a learnable matching network to capture the relationship between the novel keypoints and the super-keypoints, resulting in a more reliable matching. ESCAPE mitigates overfitting by directly transferring learned knowledge to novel categories while it does not use keypoint identifiers. We achieve state-of-the-art performance on the standard MP-100 benchmark. Our code is available at https://github.com/khoiucd/escape-tgt.

1. Introduction

2D pose estimation aims to locate the semantic keypoints of an object of interest from an input image. This task is crucial for many visual understanding tasks, such as understanding human/animal behavior or interpreting the surrounding environment in autonomous driving. Existing works focus on detecting keypoints of only one specific (super-)category, such as human, animal, or vehicle. They are inefficient in the sense that they need to train a separate model for each application. Moreover, annotating a large number of samples for some categories, e.g., rare species, is too expensive, if not impossible, and thus hinders the usefulness of the deep learning approach. A unified model that can quickly adapt to keypoints of unseen categories is desirable. Addressing this, Xu et al. [42] propose the task of category-agnostic pose estimation (CAPE) in which a unified model is trained to detect the poses of objects from any category with limited supervision.

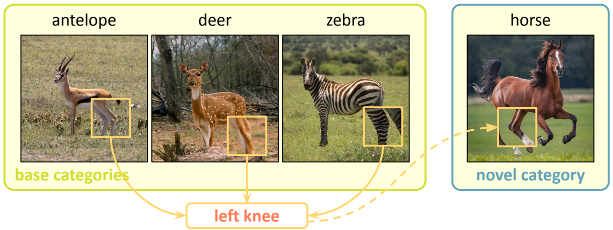

Figure 1. We encode super-keypoint 'left knee' from the 'left knee' of base animals and use the super-keypoint in adaptation to the left knee of novel animal.

Previous works attempts to compute keypoint prototypes from support and query samples and use them to locate keypoint positions on the query image via a matching step. Xu et al. [42] introduce POMNet, a transformer-based network, to capture the interactions between support and query images into the keypoint prototypes. POMNet is prone to overfitting since the keypoint prototypes heavily rely on very few, typically one or five, support samples. Recently, Shi et al. [31] propose a two-stage framework, called CapeFormer, to further calibrate the matching results. Although effective, CapeFormer uses support keypoint identifier to alleviate the ambiguity of keypoints with similar appearances. The keypoint identifier requires additional effort to consistently annotate the keypoints of different categories. This might prohibit their application in scenarios where keypoint identifiers are not available.

In this work, we aim to improve the representation capacity of the keypoint prototypes by leveraging the knowl-

*Work done as an intern at NUS.

edge learned from the base categories. We propose a Bayesian framework that models a prior over the keypoint prototypes. This allows us to analytically encode the knowledge from the base categories and explicitly transfer it to the novel ones. Specifically, we introduce the notion of super-keypoint, a high-level abstract keypoint that models the distribution of semantically related keypoints from different categories. This is based on the observation that keypoints from different categories can be closely related. For example, keypoints corresponding to the left knee of many animals are visually similar and can be grouped into a super-keypoint 'left knee'. Encoding the super-keypoint 'left knee' from many base animals can help generalize to the left knee of an unseen animal (see Fig. 1). The super-keypoints are discovered and estimated automatically from the keypoints of base categories, whose keypoint distributions are well estimated from the abundant number of samples. The adaptation to an unseen category involves two steps: first, we match each novel keypoint to a related super-keypoint; and second, we estimate the keypoint prototypes by considering both the features from the support samples and the prior knowledge encoded in the matched super-keypoints. The first step requires finding an optimal bipartite matching between the set of novel keypoints and the learned super-keypoints. To solve this, we leverage a learnable matching network to capture the structural relations between the novel keypoints and the super-keypoints before making predictions.

We call our framework **E**ncoding **S**uper-keypoints for **C**ategory-**A**gnostic **P**ose **E**stimation, dubbed ESCAPE. ESCAPE prevents the estimated keypoints from overfitting to the sparse support samples by allowing knowledge from the abundant base categories to be explicitly transferred to the novel one during adaptation. Furthermore, ESCAPE does not use keypoint identifiers to alleviate keypoint ambiguity. Experiments show that our method outperform state-of-the-art methods on the standard MP-100 benchmark, assuming keypoint identifiers are not available.

2. Related works

2D pose estimation. 2D pose estimation aims to estimate the positions of a set of semantic keypoints of an object from its image. This task holds significant importance across various applications, such as human pose estimation for understanding human behaviors [19, 24, 40, 41], animal pose estimation for wildlife conservation [11, 17, 20, 46], and more. Existing methods can be roughly divided into regression-based methods [3, 26, 34, 35, 37, 39] and heatmap-based methods [1, 2, 4, 5, 18, 33]. The former directly predicts the coordinates of keypoints, while the latter infers heatmaps and takes the peak positions as it predictions. However, these methods are primarily designed for estimating poses for a single (super-)category with a fixed set of semantic keypoints, and they typically require access to large-scale datasets.

Few-shot learning. Few-shot learning focuses on learning novel concepts from few examples [14, 45]. While few-shot learning has been extensively studied in domains such as image classification [16] and video classification [25], its application in 2D pose estimation remains relatively underexplored. Our notion of super-keypoints is related to the Bayesian approach in few-shot learning literature [29, 43, 48]. Salakhutdinov et al. [29] introduce super-categories that play role as higher-level abstract concepts for categories and encode distributions over category parameters. Zhang et al. [48] design a Bayesian framework to learn a single super-category that induces a prior over parameters for all categories; such an approach limits model flexibility and is not suitable for CAPE due to the diversity of keypoints. Yang et al. [43] observe that similar categories have nearly identical statistics and propose enhancing novel categories with similar base categories. However, it is not trivial to adopt this method for CAPE due to the lack of a structural way to measure similarity between categories with semantically and quantitatively different keypoints.

Semantic correspondence. Semantic correspondence refers to the semantic matching between pixels or keypoints between images [6]. A semantic correspondence model typically includes a feature extractor that generates dense features from images and establishes the correspondence using a similarity function. Oquab et al. [27] show that self-supervised ViT models can produce expressive features for finding semantic correspondence. Recently, Tang et al. [36] observe that semantic correspondence naturally emerges as an ability of diffusion models These models can effectively solve CAPE by matching query pixels to keypoints in support images. Since ESCAPE operates in the feature space, it complements semantic correspondence models. Integrating ESCAPE with an off-the-shelf feature extractor consistently enhances its performance in CAPE tasks.

3. Background

3.1. Category-agnostic pose estimation

Category-agnostic pose estimation (CAPE) [42] seeks an unified model to estimate poses from any category given a few annotated samples. Each category y has a set of T_y keypoints of interest: $J_y = \{j_r\}_{r=1}^{T_y}$. Given an image $\mathbf{I}$ of y, the model's objective is to accurately predict the coordinates $P = \{\mathbf{p}_r\}_{r=1}^{T_y}$ of these keypoints. The setting of CAPE consists of two disjoint sets of categories that are C_{base} for training the model and C_{novel} for testing its generalization capabilities. While the base categories have abundant training data, the novel categories have very few training samples. Specifically, a training set $\mathcal{D}^y_{base} = \{(\mathbf{I}^i, P^i)\}^{n_y}_{i=1}$ (n_y is large) of each base category y from C_{base} is available for training the model. The testing stage includes a large number of episodes (or tasks) sampled from the novel categories. In each episode, a support set $\mathcal{S}^y_{novel} = \{(\mathbf{I}^i, P^i)\}^{K}_{i=1}$ of K annotated samples from $y \in C_{novel}$ is given to the model for adaptation, and the model performance is evaluated on a set $\mathcal{Q}^y_{novel} = \{(\mathbf{I}^i, P^i)\}^{K+Q}_{i=K+1}$ of Q query samples.

Support keypoint identifier. The MP-100 dataset is composed of existing widely used datasets, many of which impose an ordering over the keypoints of their categories. Shi et al [31] refer to this ordering information as support keypoint identifier and encode it into the keypoint features to alleviate the ambiguity of novel keypoints. However, incorporating keypoint identifiers may limit the applicability of CAPE to real-world scenarios where novel keypoints can be freely defined. In this paper, we focus on the challenging yet practical settings for CAPE that do not assume the order of keypoints to be available. We refer readers to [13] for detailed discussion of the support keypoint identifier.

3.2. Simple baseline for CAPE

In this section, we introduce a simple yet effective baseline for CAPE. The baseline includes a feature extractor f_θ that extracts dense features from input images. We denote the local feature extracted at a position $\mathbf{p}$ of an input image $\mathbf{I}$ as $f_\theta(\mathbf{I})[\mathbf{p}] \in \mathbb{R}^d$. We aim to train f_θ so that it possesses strong semantic correspondence for the keypoints of interest.

For each base category y with keypoints of interest J_y, we define a set of prototypes $\mathbf{W}^y = \{\mathbf{w}_r\}^{T_y}_{r=1}$; each prototype $\mathbf{w}_r \in \mathbb{R}^d$ is responsible for predicting a heatmap $\mathbf{H}_r$ for the r^{th} keypoint in J_y as follows:

$$\mathbf{H}_r[\mathbf{p}] = \langle f_\theta(\mathbf{I})[\mathbf{p}], \mathbf{w}_r \rangle, \tag{1}$$

where $\langle \cdot\ ,\ \cdot \rangle$ denotes the inner product, and $\mathbf{p}$ varies over the spatial dimensions. The peak position in the heatmap is then the prediction for that keypoint.

We train the feature extractor f_θ and the base prototypes $\{\mathbf{W}^y\}_{y\in C_{base}}$ on the base training data $\bigcup_{y\in C_{base}}\{\mathcal{D}^y_{base}\}$ with the standard Mean Squared Error (MSE) loss [23, 38, 41] between the predicted heatmaps and the ground-truth heatmaps.

To adapt to a novel category $y \in C_{novel}$ with support and query sets $(\mathcal{S}^y_{novel}, \mathcal{Q}^y_{novel})$, we freeze the feature extractor and compute the prototypes $\mathbf{W}^y = \{\mathbf{w}_r\}^{T_y}_{r=1}$ for novel keypoints as the means of keypoint features extracted from the support set:

$$\mathbf{w}_r = \frac{1}{K} \sum_{(\mathbf{I}^i, P^i)\in \mathcal{S}^y_{novel}} f_\theta(\mathbf{I}^i)[\mathbf{p}^i_r]. \tag{2}$$

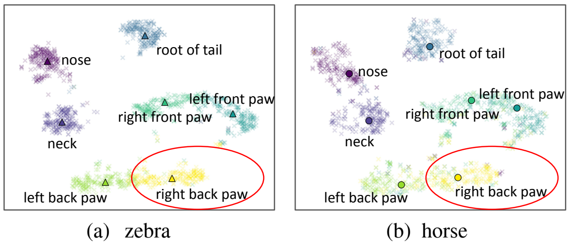

Figure 2. t-SNE visualization of distributions of 7 keypoints of zebra and horse. Different colors represent different keypoints. '×' denotes keypoint features extracted from images. (a) shows extracted features (denoted as '×') and distribution means (denoted as '△') for keypoints of **zebra**. (b) shows extracted features (denoted as '×') and distribution means (denoted as '○') for keypoints of **horse**.

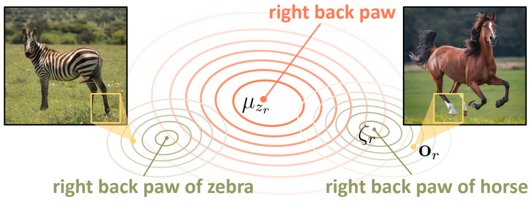

Figure 3. Super-keypoint modeling. Keypoint 'right back paw of zebra' and keypoint 'right back paw of horse' belong to the same super-keypoint 'right back paw'. The keypoints distribute closely around their super-keypoint.

The computed prototypes are then used to predict keypoint heatmaps for query images as in Eq. 1.

Our simple baseline provides results comparable to POMNet [42], which has more complicated matching network. The baseline, however, has the same problem of overfitting as POMNet and CapeFormer due to its heavy reliance on the sparse support samples. Building on our simple baseline, we propose a Bayesian framework that allows us to estimate better prototypes for the novel keypoints, while keeping the feature extractor the same. Specifically, we seek to equip the baseline with strong a prior to prevent the keypoint prototypes from overfitting to the support set.

4. Prior as a mixture of super-keypoints

We notice that keypoints of different categories can be visually similar, and hence their features should follow the same distribution. To verify our hypothesis, we use the feature extractor from our baseline model to extract features at the positions of seven keypoints in images of 'zebra' (a base category) and 'horse' (an unseen category). Fig. 2 (a) and (b) show keypoints features extracted from zebra and horse images respectively. For each keypoint, we also visualize the mean of its distribution (denoted as △ for keypoint of zebra and ○ for keypoint of horse). As shown in the figure, although the model has not seen any horse sample during training, the keypoint features of horse still align well with those of zebra. This alignment suggests that capturing the distributions of zebra keypoints facilitates generalization to horse keypoints. From this observation, we introduce super-keypoints to model the connections between keypoints of horse and zebra. For instance, the keypoint 'right back paw of zebra' and the keypoint 'right back paw of horse' belong to the same super-keypoint 'right back paw'.

4.1. Super-keypoints modeling

We introduce the notion of super-keypoint to define a prior over a group of visually similar keypoints from different categories. Let us consider the r^{th} keypoint of a category y. Given an image $\mathbf{I}$ of y, we denote $\mathbf{o}_r := f_\theta(\mathbf{I})[\mathbf{p}_r]$ as the local feature extracted at the position of the r^{th} keypoint in $\mathbf{I}$. We assume that the keypoint feature $\mathbf{o}_r$ follows a Gaussian distribution with a mean ζ_r and covariance $\phi^2 I$. We further assume that the mean of the keypoint distribution ζ_r obeys a Gaussian distribution with a mean μ_{z_r} and covariance $\sigma^2 I$, where μ_{z_r} represent the super-keypoint of the r^{th} keypoint. Formally,

$$\mathbf{o}_r \sim \mathcal{N}(\mathbf{o}|\zeta_r; \phi^2), \quad \text{and} \quad \zeta_r \sim \mathcal{N}(\zeta|\mu_{z_r}; \sigma^2). \tag{3}$$

Fig. 3 illustrates our super-keypoint modeling. Intuitively, μ is the representation of a super-keypoint in the feature space, and its keypoint members distribute closely nearby. Our goal is to discover and estimate a set of super-keypoints from keypoints of base categories so that the super-keypoints can encode the statistics of groups of related keypoints across different categories. When adapting to a novel category, the statistics encoded in the super-keypoints are directly transferred to the prototypes of novel keypoints, preventing our model from overfitting to the sparse support set.

4.2. Super-keypoints discovering and learning

We obtain the set of super-keypoints by applying a clustering method to the keypoints of all base categories. Specifically, given the trained feature extractor f_θ from the baseline, we extract the features of keypoints from all base categories. We then estimate the distribution mean of each keypoint in the base categories as follows:

$$\zeta_r^y = \frac{1}{|\mathcal{D}_{base}^y|} \sum_{(\mathbf{I}^i, P^i) \in \mathcal{D}_{base}^y} f_\theta(\mathbf{I}^i)[\mathbf{p}_r^i], \tag{4}$$

where ζ_r^y represents the distribution mean of the r^{th} keypoint of category $y \in C_{base}$. The set of keypoints from all base categories is then $\bigcup_{y \in C_{base}} \{\zeta_r^y\}_{r=1}^{T_y}$. In this work, we use the k-nearest neighbor based density peaks clustering [9, 47], which does not require the number of clusters to be known in advance. Note that keypoints of the same category are distinct, so they cannot belong to the same super-keypoint. A detailed description of the clustering is provided in the Appendix. Finally, we estimate each super-keypoint representation μ as the empirical mean of distribution means of keypoints within the cluster, resulting in a set of L super-keypoints: $\mathbf{M} = \{\mu_1, \mu_2, \ldots, \mu_L\}$.

5. Novel category adaptation

In this section, we describe how to adapt to a novel category y whose support and query sets are $(\mathcal{S}_{novel}^y, \mathcal{Q}_{novel}^y)$. Our goal is to estimate the prototypes that well represent the distributions of the novel keypoints from a few support samples. In particular, for the r^{th} keypoint of y, we want the estimated prototype $\mathbf{w}_r$ to be close to the true distribution mean ζ_r as much as possible. With this interpretation, the estimated prototype for a novel keypoint in the baseline (Eq. 2) can be viewed as maximum likelihood estimate of the mean of the keypoint distribution. It is clear that the baseline is prone to overfitting since the number of support samples is extremely limited.

5.1. Prototype estimate via maximum a posteriori

We propose to alleviate the overfitting by letting the statistics captured in the super-keypoints to be transferred to the novel keypoints. Specifically, consider the r^{th} keypoint of y, we compute its prototype $\mathbf{w}_r$ as maximum a posterior estimate of ζ_r:

$$\mathbf{w}_r^{MAP} = \underset{\zeta_r}{\text{argmax}} \ \log p(\zeta_r | \{\mathbf{o}_r^i\}_{i=1}^K, \mathbf{M}), \tag{5}$$

where $\mathbf{o}_r^i = f_\theta(\mathbf{I}^i)[\mathbf{p}_r^i]$ is the feature of the r^{th} keypoint extracted from the i^{th} support sample in $\mathcal{S}_{novel}^y$, and $\mathbf{M} = \{\mu_1, \mu_2, \ldots, \mu_L\}$ is the set of learned super-keypoints.

However, when adapting to a novel category, the super-keypoints of the novel keypoints are not available in general. We do not know which super-keypoint the r^{th} novel keypoint should belong to. To this extent, we adopt the expectation maximization (EM) algorithm [21], which first matches the r^{th} keypoint with the 'best' super-keypoint from $\mathbf{M}$ (the expectation step) and then optimizes for the prototype with the prior encoded in the matched super-keypoint (the maximization step). Below, we summarize the two steps; detailed derivation of the (EM) algorithm is provided in the Appendix.

Expectation step. Let z_r be the super-keypoint assignment of the r^{th} keypoint of y. We align the keypoint with the most likely super-keypoint $z_r^\star$, i.e, the z_r that maximizes $\log p(\zeta_r, z_r | \{\mathbf{o}_r^i\}_{i=1}^K, \mathbf{M})$. Note that keypoints of the same category cannot belong to the same super-keypoint. Therefore, we further adopt the Hungarian algorithm [15]

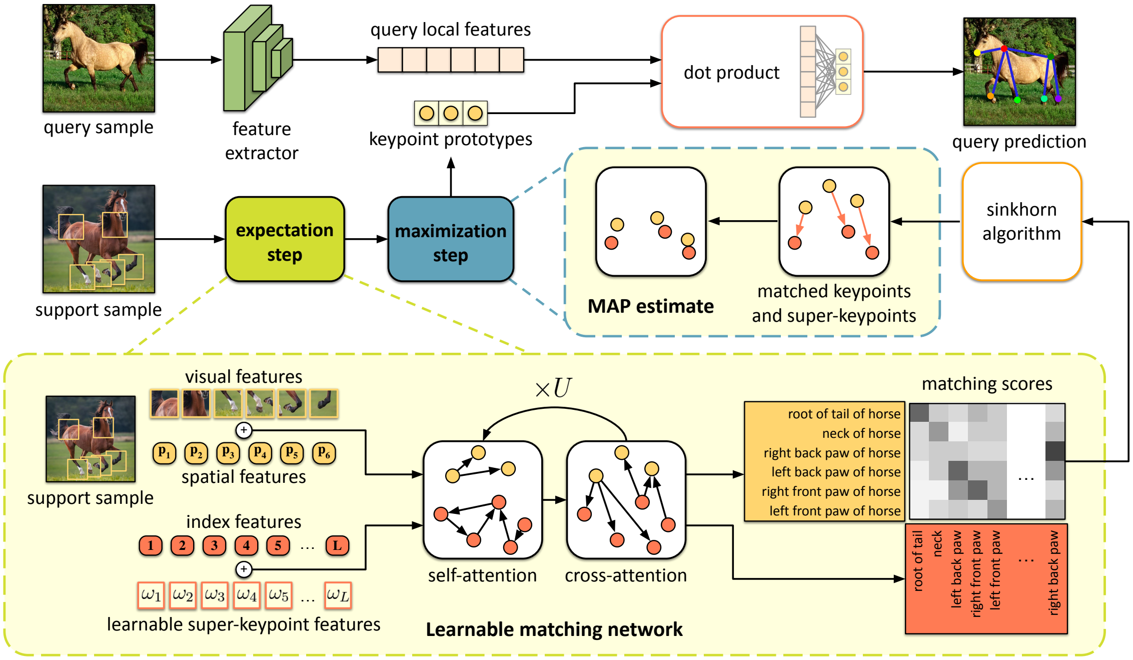

Figure 4. Illustration of our adaptation procedure for a novel category. We leverage the EM algorithm to probabilistically optimize for the keypoint prototypes, considering both the support samples and the learned super-keypoints. We first align the novel keypoints with the different super-keypoints via a learnable matching network (expectation step). We then let the statistics encoded in the matched super-keypoints to be transferred directly to the keypoint prototypes (maximization step). The keypoint prototypes can be used to estimate poses for query samples.

to optimally align the T_y keypoints of y with distinct super-keypoints $\{\mu_{z_r^\star}\}_{r=1}^{T_y} \subset \mathbf{M}$. We refer to this matching procedure as **simple matching** for the rest of the paper.

Maximization step. Now, we know the 'best' super-keypoint for each keypoint of y. We can compute their prototypes analytically as follows:

$$\mathbf{w}_r^{MAP} = \underset{\mathbf{w}_r}{\operatorname{argmax}} \log p(\mathbf{w}_r | \{\mathbf{o}_r^i\}_{i=1}^K, \mu_{z_r^\star}) \qquad (6)$$

$$= \alpha \frac{1}{K} \sum_{i=1}^{K} \mathbf{o}_r^i + (1-\alpha)\mu_{z_r^\star}, \qquad (7)$$

where $\alpha \in (0,1)$ depends on ϕ and σ. Since the distribution mean of a keypoint must distribute near its super-keypoint, the second term in Eq. 7 has an effect of pulling the estimated prototype closer to the true mean of the keypoint distribution, diminishing the dependence of the prototype on the support samples.

5.2. Learnable matching network

Matching novel keypoints with semantically related super-keypoints is crucial for the success of our method. A correct super-keypoint can pull the estimated prototype closer to the distribution mean of the keypoint under consideration, while an incorrect super-keypoint pulls the prototype away. However, as later shown in the experiments, the simple matching in the expectation step fails to produce reliable super-keypoints assignments. We argue that this is because the simple matching treats the novel keypoints independently, discarding their structural relations. It ignores the keypoint coordinates in the support samples, and so the model cannot reason about poses of the object of interest.

To overcome these limitations, we replace the simple matching in the expectation step with a learnable matching network to simultaneously produce matching scores between keypoints in an image and the learned super-keypoints. The design is inspired by SuperGlue [30], a matching framework that has shown great performance in aligning two sequences of features.

Input features. Our matching network takes two sequences of features as input and produces the matching score matrix for their elements. The first sequence is augmented features of T_y keypoints extracted from a support image of y. The second sequence is a set of features for the L super-keypoints. To encode features for the keypoints, we consider not only their visual appearance but also their coordinates within the support image. Specifically, given

Method	Split 1	Split 2	Split 3	Split 4	Split 5	Mean PCK	Split 1	Split 2	Split 3	Split 4	Split 5	Mean PCK
ProtoNet [32]	46.05	40.84	49.13	43.34	44.54	44.78	60.31	53.51	61.92	58.44	58.61	58.56
MAML [10]	68.14	54.72	64.19	63.24	57.20	61.50	70.03	55.98	63.21	64.79	58.47	62.50
Finetune [22]	70.60	57.04	66.06	65.00	59.20	63.58	71.67	57.84	66.76	66.53	60.24	64.61
POMNet [42]	84.23	78.25	78.17	78.68	79.17	79.70	84.72	79.61	78.00	80.38	80.85	80.71
CapeFormer [31]	84.13	79.90	79.58	79.78	80.34	80.75	90.46	86.47	**85.80**	86.47	**86.71**	87.19
ESCAPE	**86.89**	**82.55**	**81.25**	**81.72**	**81.32**	**82.74**	**91.41**	**87.43**	85.33	**87.27**	**86.76**	**87.63**

Table 1. Comparison with state-of-the-art works on MP-100 dataset under 1-shot and 5-shot settings.† denotes keypoint identifiers.

support image $\mathbf{I} \sim \mathcal{S}^y_{novel}$, the keypoint features are:

$$\mathbf{u}_r = f_\gamma(\mathbf{I})[\mathbf{p}_r] + PE_{\text{geo}}(\mathbf{p}_r), \quad \text{for } r = \{1, \dots, T_y\}, \quad (8)$$

where f_γ is a CNN feature extractor, and PE_{geo} embeds a coordinate into the feature space. By considering the keypoint coordinates, our model can reason about the pose of the object in the support image.

At the other end, the features for the super-keypoints are defined as:

$$\mathbf{v}_z = \omega_z + PE_{\text{abs}}(z), \qquad \text{for } z \in \{1, \dots, L\}, \qquad (9)$$

where ω_z is a learnable feature vector for the z^{th} super-keypoint, and PE_{abs} embeds an index $(1, \dots, L)$ into the feature space. Note that the indexes of the super-keypoints are obtained from the clustering step described in Sec. 4.2.

Feature aggregation with attention. The two sequences of features $\{\mathbf{u}_r\}_{r=1}^{T_y}$ and $\{\mathbf{v}_z\}_{z=1}^{L}$ are then passed to a neural network for information aggregation. The network comprises U blocks of self-attention and cross-attention to model interactions both within and between the elements of the sequences. Results are two sequences of refined features: $\{\mathbf{u}^\star_r\}_{r=1}^{T_y}$ and $\{\mathbf{v}^\star_z\}_{z=1}^{L}$. The predicted matching scores matrix is then $\mathbf{S} \in \mathbb{R}^{T_y \times L}$, with each element $\mathbf{S}_{r,z} = \langle \mathbf{u}^\star_r, \mathbf{v}^\star_z \rangle$ measuring the matching score between the r^{th} keypoint and the z^{th} super-keypoint. An illustration of the matching network is provided at the bottom of Fig. 4.

Matching as optimal transport problem. Since the keypoints within the same category must be matched to different super-keypoints, our goal is to find a maximum score bipartite matching between the two sequences. Following SuperGlue [30], we replace the Hungarian algorithm with the differentiable Sinkhorn algorithm [7] for optimal transport. It takes the matching score matrix $\mathbf{S}$ as input and produces a matrix $\mathbf{Q} \in \mathbb{R}^{T_y \times L}$ where each element $\mathbf{Q}_{r,z}$ represents the probability that the r^{th} keypoint is matched with the z^{th} super-keypoint. We train the network end-to-end on the keypoints of base categories, whose super-keypoints assignments are available from the clustering step. Details of the Sinkhorn algorithm and the training procedure of the matching network are provided in the Appendix.

Incorporating learnable matching network to novel category adaptation. We integrate the learned matching network into the novel category adaptation. Specifically, for every support image $\mathbf{I}^i \in \mathcal{S}^y_{novel}$, we compute the matching score matrix $\mathbf{S}^i$ from its keypoints and the learned super-keypoints. The corresponding optimal assignment matrix is then denoted as $\mathbf{Q}^i$. The final prototype for the r^{th} keypoint are then estimated as follows:

$$\mathbf{w}_r^{\text{ESCAPE}} = \alpha \frac{1}{K} \sum_{i=1}^{K} \mathbf{o}_r^i + (1-\alpha) \frac{L}{K} \sum_{i=1}^{K} \sum_{z=1}^{L} \mathbf{Q}^i_{r,z} \mu_z. \qquad (10)$$

Here we keep the soft assignment form of $\mathbf{Q}^i$ instead of the hard assignment to model uncertainty. Our adaptation procedure is shown in Fig. 4.

6. Experiments

We perform experiments on Multi-category Pose (MP-100) proposed in [42]. The dataset consists of more than 20,000 data instances of 100 categories from 8 super-categories including human hand, human face, human body, animal body, animal face, clothes, furniture, and vehicle. We also follow the five splits from [42] for standard CAPE settings. Each split divides the dataset into 70 training categories, 10 validation categories, and 20 testing categories. The training set is used as base categories, while validation and testing sets are novel categories for validation and testing, respectively. Please refer to [42] for more details.

6.1. Experimental details

Unless otherwise specified, for fair comparisons with previous works, we utilize ResNet-50 [12] pretrained on ImageNet [8] as the feature extractor. Also following [42], we crop the object of interest from the original image using the ground-truth bounding box, and resize it to a fixed size of 256 × 256.

Probability of correct keypoint [44] with a threshold of 0.2 (PCK@0.2) is the evaluation metric for MP-100. During testing, we generate 60,000 episodes evenly distributed across the test categories and report the average PCK@0.2 over these episodes.

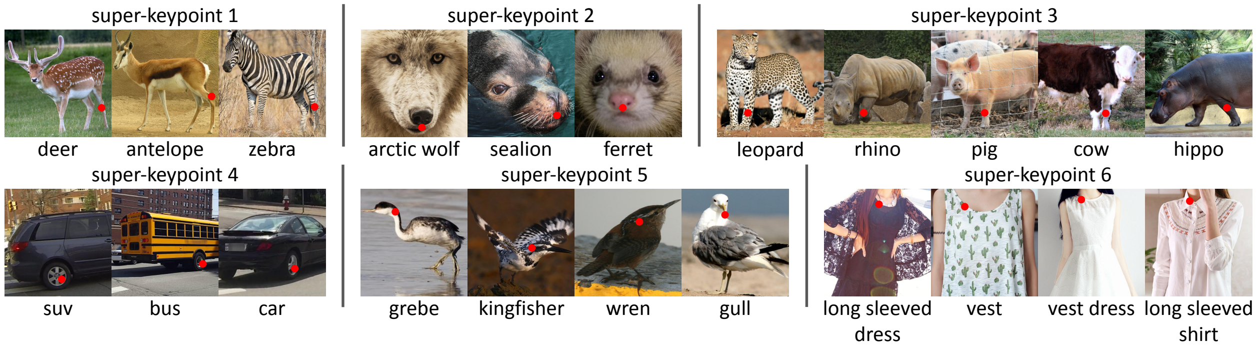

Figure 5. Keypoints of four different super-keypoints discovered from the base categories. Below each keypoint is the name of its category.

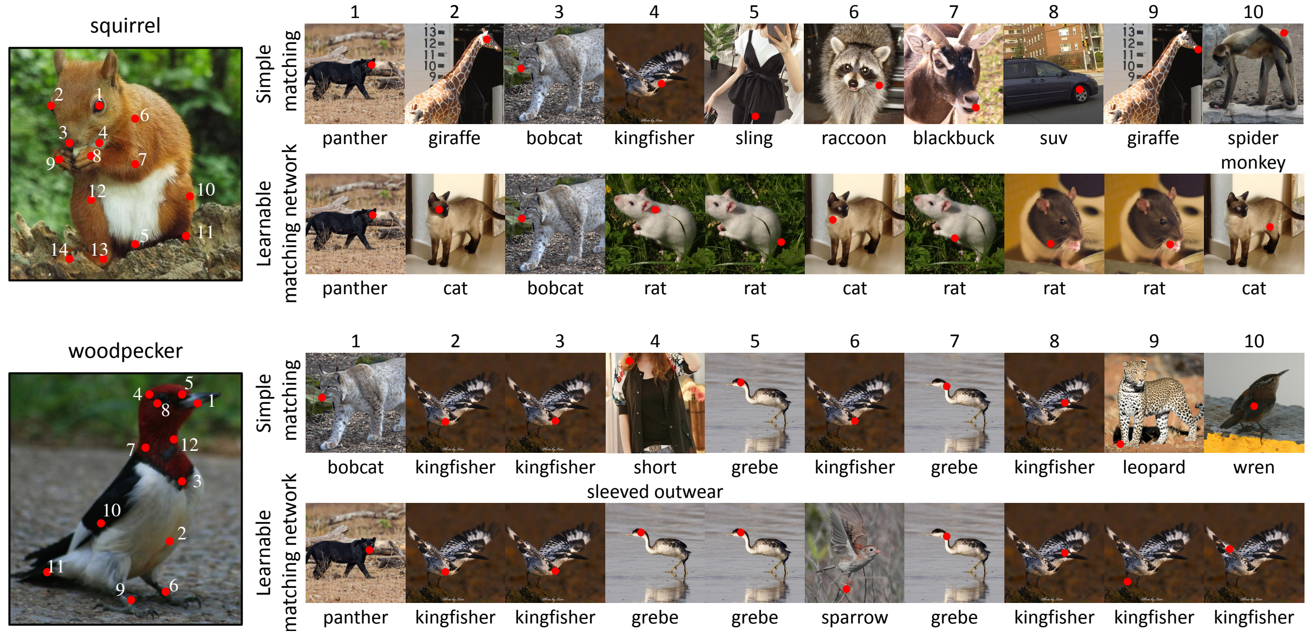

Figure 6. Qualitative comparison between simple matching and learnable matching network in aligning novel keypoints with super-keypoints. We show one base keypoint member of each matched super-keypoint.

6.2. Comparison with previous methods

We compare ESCAPE against previous methods, including Prototypical Networks [32], MAML [10], Finetune [22], POMNet [32] and CapeFormer [31]. Note that we remove the support keypoint identifier from CapeFormer. The results are shown in Tab. 1. We can see that ESCAPE outperforms other methods by around 2.5% and 0.5% in 1-shot and 5-shot settings, respectively, establishing a new state-of-the-art on the MP-100 benchmark.

6.3. Ablation study

Super-keypoints discovering. We perform clustering on the keypoints of base categories to form super-keypoints. To verify that we can obtain meaningful super-keypoints from the clustering process, we show some keypoints members of a few super-keypoints in Fig. 5. We see that each super-keypoint represents a meaningful concept. For example, super-keypoint 1 consists of ‘left knee’ of *deer*, *antelope* and *zebra*. super-keypoint 4 consists of ‘right rear wheel’ of *suv*, *bus*, and *car*, while super-keypoint 6 is a landmark on the collars of different clothes categories.

Novel keypoints and super-keypoints matching. We qualitatively compare the performance of the simple matching and the learnable matching network in aligning novel keypoints with the learned super-keypoints. Fig. 6 shows the matching results for two novel categories, namely *squirrel* and *woodpecker*. The left side of the figure shows support images of the categories, while the right side shows the super-keypoints aligned by the simple matching and the learnable matching network. For each super-keypoint, we only show one base keypoint member. Compared with the

Method	PCK
MLE	83.00 ± 0.16
MAP	83.88 ± 0.15
ESCAPE	$\mathbf{86.89 \pm 0.14}$

Table 2. Results of different prototype estimation methods, namely MLE, MAP and ESCAPE.

Feature extractor	Method	PCK
DIFT [36]	MLE	60.61
DIFT [36]	**ESCAPE**	62.47
DINOv2 [27]	MLE	88.31
DINOv2 [27]	**ESCAPE**	89.14

Table 3. Results of integrating ESCAPE to feature extractors of different semantic correspondence models

simple matching, our learnable matching network provides more reliable super-keypoints for the novel keypoints. In particular, the simple matching aligns keypoints of squirrel with super-keypoints of keypoints of kingfisher, sling and suv, which are irrelevant to the squirrel. On the other hand, the learnable matching network aligns the novel keypoints with their actual concepts, e.g., the 'left elbow of squirrel' is matched with the super-keypoint of the 'left elbow of rat'.

Ablation on novel category adaptation. In Tab. 2, we study the performance of different methods to estimate keypoint prototypes for a novel category, namely MLE, MAP and ESCAPE. MLE is our baseline, which estimate keypoint prototypes as means of support features as in Eq. 2. MAP and ESCAPE compute keypoint prototypes as maximum a posterior estimate with the EM algorithm. The difference between MAP and ESCAPE is that MAP (Eq. 7) uses the simple matching in the expectation step, whereas ESCAPE (Eq. 10) leverages the learnable matching network. We observe that incorporating super-keypoints into the estimation of keypoint prototypes helps improve the overall performance. Specifically, the MAP outperforms MLE by roughly 1%. Regardings ESCAPE, better novel keypoints and super-keypoints alignments from the learnable matching network immediately translate into 3% improvements over the MAP counterpart.

Ablation on different feature extractors. Feature extractors from semantic correspondence models can effectively solve CAPE since they can generate dense features for support and query images. In Tab. 3, we study the compatibility of ESCAPE with feature extractors from different semantic correspondence models, namely DIFT [36] and DINOv2 [27]. We take the pretrained stable diffusion 2.1 [28] as the backbone for DIFT, whereas we use ViT-ViT-S/14 for DINOv2. As seen, ESCAPE consistently boosts the performance of the considered semantic correspondence models. This shows that ESCAPE can benefit from future research for better representation.

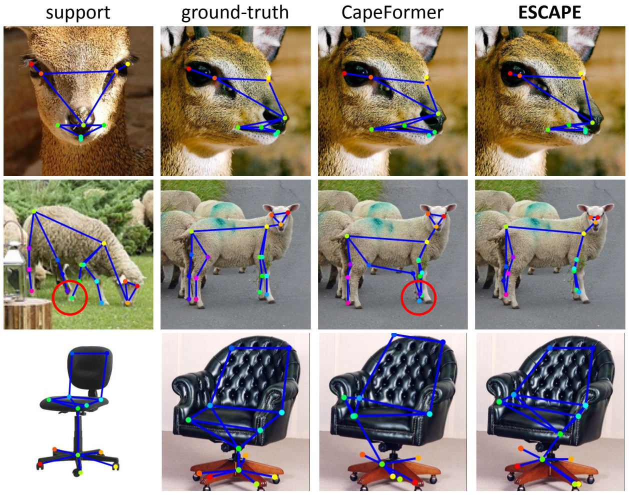

Figure 7. We qualitatively compare pose estimation of CapeFormer and ESCAPE under 1-shot setting.

6.4. Qualitative results

Figure 7 compares the pose estimation performance of CapeFormer and ESCAPE. Overall, ESCAPE gives more accurate pose predictions compared to CapeFormer. In particular, CapeFormer is overfitting to the support image of the sheep category. The predicted positions of the left front and back paws by CapeFormer are geometrically close to each other as they appear in the support image. In contrast, since ESCAPE allows learned knowledge about base animals to be transferred to the novel keypoints, it recovers the pose of the query sheep beautifully.

7. Conclusion

In this paper, we address the task of category-agnostic pose estimation by introducing ESCAPE, a Bayesian framework that induces a prior over the keypoint features of different categories. This prior is a mixture of super-keypoints, with each super-keypoint encoding a distribution of a group of related keypoints. ESCAPE enables the transfer of knowledge from base categories to the novel ones, thereby preventing the query predictions from overfitting to sparse support samples. ESCAPE offers several advantages over previous works: it exhibits less overfitting to support samples, does not rely on support keypoint identifiers, and is model-agnostic.

Acknowledgement

This research is supported in part by the National Research Foundation, Singapore under its Competitive Research Program Award NRF-CRP23-2019-0003.

References

[1] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xiangyu Zhang, Xinyu Zhou, Erjin Zhou, and Jian Sun. Learning delicate local representations for multi-person pose estimation. In *European Conference on Computer Vision*, pages 455–472. Springer, 2020. 2

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[3] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016. 2

[4] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 2

[5] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 2

[6] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *Advances in neural information processing systems*, 29, 2016. 2

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 6

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[9] Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99: 135–145, 2016. 4

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 6, 7

[11] Jacob M Graving, Daniel Chae, Hemal Naik, Liang Li, Benjamin Koger, Blair R Costelloe, and Iain D Couzin. Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife*, 8:e47994, 2019. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[13] Or Hirschorn and Shai Avidan. Pose anything: A graph-based approach for category-agnostic pose estimation. *arXiv preprint arXiv:2311.17891*, 2023. 3

[14] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021. 2

[15] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[16] Duong Le, Khoi Duc Nguyen, Khoi Nguyen, Quoc-Huy Tran, Rang Nguyen, and Binh-Son Hua. Poodle: Improving few-shot learning via penalizing out-of-distribution samples. *Advances in Neural Information Processing Systems*, 34:23942–23955, 2021. 2

[17] Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. Atrw: a benchmark for amur tiger re-identification in the wild. *arXiv preprint arXiv:1906.05586*, 2019. 2

[18] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019. 2

[19] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Changhu Wang, and Jiashi Feng. Improving convolutional networks with self-calibrated convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10096–10105, 2020. 2

[20] Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1859–1868, 2021. 2

[21] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. 4

[22] Akihiro Nakamura and Tatsuya Harada. Revisiting fine-tuning for few-shot learning. *arXiv preprint arXiv:1910.00216*, 2019. 6, 7

[23] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016. 3

[24] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017. 2

[25] Khoi D Nguyen, Quoc-Huy Tran, Khoi Nguyen, Binh-Son Hua, and Rang Nguyen. Inductive and transductive few-shot video classification via appearance and temporal alignments. In *European Conference on Computer Vision*, pages 471–487. Springer, 2022. 2

[26] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6951–6960, 2019. 2

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 8

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 8

[29] Ruslan Salakhutdinov, Joshua Tenenbaum, and Antonio Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 195–206. JMLR Workshop and Conference Proceedings, 2012. 2

[30] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 5, 6

[31] Min Shi, Zihao Huang, Xianzheng Ma, Xiaowei Hu, and Zhiguo 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*, pages 7308–7317, 2023. 1, 3, 6, 7

[32] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 6, 7

[33] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[34] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017. 2

[35] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 529–545, 2018. 2

[36] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 2, 8

[37] Zhi Tian, Hao Chen, and Chunhua Shen. Directpose: Direct end-to-end multi-person pose estimation. *arXiv preprint arXiv:1911.07451*, 2019. 2

[38] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27, 2014. 3

[39] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 2

[40] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 2

[41] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 2, 3

[42] Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Pose for everything: Towards category-agnostic pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 398–416. Springer, 2022. 1, 2, 3, 6

[43] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. *arXiv preprint arXiv:2101.06395*, 2021. 2

[44] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012. 6

[45] Wang Yaqing, Yao Quanming, T Kwok James, and M Ni Lionel. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys*, 53(3):1–34, 2020. 2

[46] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*, 2021. 2

[47] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022. 4

[48] Xueting Zhang, Debin Meng, Henry Gouk, and Timothy M Hospedales. Shallow bayesian meta learning for real-world few-shot recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 651–660, 2021. 2