Occlusion-Resilient Category-Agnostic Pose Estimation with Conditional Flow Matching

Jiyong Rao[1], Shengjie Zhao[1†], Yu Wang[1], and Hao Deng[1]

[1] School of Computer Science and Technology, Tongji University, China
shengjiezhao@tongji.edu.cn

Abstract. Category-Agnostic Pose Estimation (CAPE) aims to localize category-specific keypoints on a query image from only a few annotated support examples. Existing CAPE methods are largely based on direct coordinate regression or local heatmap matching, which become unreliable under severe occlusion because missing keypoints cannot be recovered from local evidence alone. We present **FlowCape**, a conditional flow-matching framework for occlusion-resilient CAPE. FlowCape first extracts query-image features and combines them with keypoint-description embeddings in a *Heatmap-Guided Initialization* module, producing a stable initial pose for transport. A shared *Riemannian Pose Head* then predicts the conditional velocity field, while its internal *Graph Flow Encoder* fuses query image, textual keypoint semantics, and skeleton topology to enforce structured pose evolution. The final prediction is obtained by probability-flow ODE rollout, and training is driven by hybrid flow-matching supervision together with initialization and rollout consistency constraints. To better evaluate robustness under severe occlusion, we further introduce **Occ80**, an occlusion-focused benchmark spanning 80 categories. Experiments on MP-100 and Occ80 show that FlowCape consistently outperforms strong CAPE baselines, reach the state-of-the-art under occluded scenario.

Keywords: Category-Agnostic Pose Estimation · Flow Matching · Few Shot Learning

1 Introduction

Category-Agnostic Pose Estimation (CAPE) [8,16,25,28,29,33] aims to localize category-defined keypoints on a query image given only a few annotated support examples. In contrast to category-specific pose estimation [26, 31, 34–36], where training and evaluation are confined to specific categories with fixed skeleton, CAPE requires models to transfer keypoint knowledge to previously unseen categories whose shapes, part definitions, and topological structures may differ substantially. This distinction becomes especially critical under severe occlusion: once local visual evidence is missing, category-specific models can often rely on class-specific appearance and pose priors learned from closed-set supervision,

† Corresponding author.

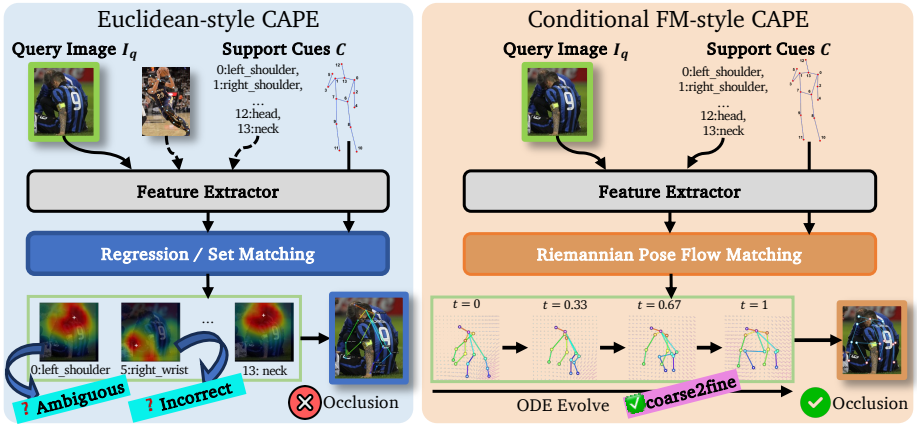

Fig. 1: Conceptual comparison between Euclidean-style CAPE and conditional FM style CAPE. Conventional CAPE performs discriminative query–support feature matching to directly predict keypoints in Euclidean space, which may produce over-confident predictions under occlusion. In contrast, conditional FM-style models CAPE as conditional generative flow on a pose manifold, progressively transforming initial states into occlusion-consistent configurations through Riemannian flow matching.

whereas CAPE must infer invisible keypoints for novel categories without such category-dependent regularities. As a result, occlusion in CAPE is not merely a missing-evidence problem, but a coupled challenge of correspondence ambiguity and structure-aware reasoning, requiring the model to recover occluded keypoints while preserving semantic consistency and topological plausibility.

Most existing CAPE methods remain fundamentally one-shot and discriminative. They usually predict coordinates directly or regress local heatmaps in $\mathbb{R}^2$ (shown in Fig. 1), which is effective when keypoints are visible but brittle when visual evidence is missing. Two issues follow. First, local prediction alone provides weak global structural control, so heavily occluded keypoints often drift to implausible configurations. Second, the model is still trained to output a single deterministic estimate, which tends to produce over-confident predictions even when the observation is ambiguous. Robust CAPE under occlusion therefore requires a formulation that can gradually refine pose hypotheses while preserving semantic and structural consistency.

Flow Matching [17, 19] offers a natural alternative because it models prediction as continuous transport from an initial state to a target pose through a learned conditional velocity field. However, naive applications of Flow Matching [5, 13, 32] is not sufficient for CAPE. The transport must start from a meaningful initial pose rather than an uninformed prior, and the velocity field must jointly reason over query appearance, keypoint semantics, and skeleton structure throughout the rollout. Otherwise, the trajectory is easily destabilized by poor initialization or by a weakly constrained vector field.

To address these issues, we propose **FlowCape**, a conditional flow-matching framework tailored to occlusion-resilient CAPE. Given a query image, keypoint-description embeddings, and a skeleton graph, FlowCape first extracts query features with a Image Encoder and feeds them to a *Heatmap-Guided Initialization* module. This module converts the interaction between query features and textual keypoint semantics into per-keypoint similarity maps and a stable initial pose. Starting from this initialization, a shared *Riemannian Pose Head* predicts the transport velocity. Inside the head, a *Graph Flow Encoder* updates keypoint tokens by combining query image memory, textual semantics, and graph structure, so that the learned flow remains both appearance-aware and topology-aware. The final pose is then obtained by a probability-flow ODE rollout that repeatedly applies the same head along the transport trajectory. During training, we supervise this process with hybrid flow-matching losses under a skeleton-aware Laplacian metric, together with initialization and rollout consistency constraints.

Although occlusion is a defining challenge in CAPE, existing benchmarks such as MP-100 do not explicitly emphasize severe occlusion evaluation. We therefore introduce **Occ80**, an occlusion-focused benchmark spanning 80 categories and curated severe-occlusion samples. Comprehensive experiments on MP-100 and Occ80 demonstrate that FlowCape consistently outperforms strong CAPE baselines, with particularly large gains when accurate localization depends on both semantic correspondence and global structural reasoning. The main contributions of this work are summarized as follows:

- We propose FlowCape, a conditional flow-matching framework that formulates CAPE as structured pose transport and is explicitly designed for occlusion-resilient estimation.
- We design **Heatmap-Guided Initialization** mechanism and a shared **Riemannian Pose Head** to jointly integrate query-image evidence, keypoint-description semantics, and skeleton topology across pose initialization and flow prediction.
- We develop flow-matching supervision under a skeleton-aware Laplacian metric, together with initialization and rollout consistency constraints, to improve structural recovery for occluded keypoints.
- We establish **Occ80**, an occlusion-focused benchmark for evaluating robustness and cross-category generalization in CAPE.

2 Related Works

2.1 Category-Agnostic Pose Estimation

Category-Agnostic Pose Estimation was pioneered by POMNet [33], which formulated the task as a keypoint matching problem between a support image and a query image. Subsequent research, such as CapeFormer [29], extended this paradigm into a two-stage DETR-like framework to refine initial matching results. To better handle structural consistency and occlusions, GraphCape [8]

introduced a graph-based representation, treating keypoints as connected nodes rather than isolated entities. This was further refined by GenCape [25], which adaptively learns the adjacency matrix to capture varying structural dependencies between joints.

Meanwhile, the field has moved towards more flexible supervision signals. CapeX [28] introduced textual point explanations, replacing the need for a support image with semantic descriptions and skeletal pose-graphs. This trend culminated in CapeLLM [12], the first multimodal large language model (MLLM [6, 20]) for CAPE, which leverages advanced language priors to estimate keypoints for unseen categories in a support-free manner. However, most existing CAPE methods remain deterministic regression models that struggle with the "generative fallacy" under severe occlusion. While diffusion-based models like DiffPose [11,30] address uncertainty, they suffer from high computational costs. Our work fills this gap by utilizing the efficient generative capability of Flow Matching while incorporating structural rigor via Riemannian manifolds.

2.2 Flow Matching and Structure-Aware Generative Transport

Flow Matching (FM) [7,17,18] has recently emerged as a powerful alternative to diffusion models [4, 10, 30] for generative tasks. Introduced by Lipman [17], FM learns a deterministic velocity field that transports a simple prior distribution to a target data distribution via an Ordinary Differential Equation (ODE). Compared to diffusion models, FM provides simulation-free training and typically requires fewer integration steps during inference, making it highly suitable for real-time applications like depth estimation and 3D pose lifting.

To account for geometric constraints in non-Euclidean domains, Riemannian Flow Matching (RFM) [2] was introduced. RFM has been successfully applied to rigid body orientations on the SE(3) manifold [21], material crystal structures, and robotic motion policies. However, applying RFM to CAPE poses a unique challenge: the underlying manifold is not fixed but depends on the skeletal topology of the novel category. Furthermore, standard FM suffers from "coupling issues" where large distances between the prior and target result in integration drift. Our FlowCape framework addresses these by defining a category-agnostic metric via the Skeleton Laplacian and introducing a heatmap-guided initialization to ensure short, stable transport paths.

3 Preliminaries

We briefly review the conditional Flow Matching (FM) formulation. Let $\mathbf{x}_0 \in \mathbb{R}^{N\times 2}$ denote an initial pose state, $\mathbf{x}_1 \in \mathbb{R}^{N\times 2}$ denote a target pose (N is the number of keypoints), and $\mathbf{c}$ denote the condition variables. A conditional flow model with learnable parameters θ learns a time-dependent velocity field $\mathbf{v}_\theta(\mathbf{x}, t \mid \mathbf{c})$ that transports $\mathbf{x}_0$ to $\mathbf{x}_1$ through the probability flow Ordinary Differential Equations (ODE)

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{v}_\theta(\mathbf{x}(t), t \mid \mathbf{c}), \qquad \mathbf{x}(0) = \mathbf{x}_0, \tag{1}$$

where $t \in [0,1]$ is the continuous time variable. FM defines a reference path $\psi_t(\mathbf{x}_0, \mathbf{x}_1)$ between the endpoints and regresses the model velocity to the path tangent $\mathbf{x}_t = \psi_t(\mathbf{x}_0, \mathbf{x}_1)$,$\mathbf{u}_t = \partial_t \psi_t(\mathbf{x}_0, \mathbf{x}_1)$. The conditional FM objective is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{(\mathbf{x}_0, \mathbf{x}_1, \mathbf{c}),\, t} \left[\| \mathbf{v}_\theta(\mathbf{x}_t, t \mid \mathbf{c}) - \mathbf{u}_t \|_2^2 \right]. \tag{2}$$

We adopt the linear probability path $\psi_t(\mathbf{x}_0, \mathbf{x}_1) = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$. It leads to straight-line transport in Euclidean pose space while allowing the learned conditional field to absorb appearance and structure cues through $\mathbf{c}$.

At inference time, the learned ODE in Eq. (1) is solved numerically. Using explicit Euler integration with K steps and $\Delta t = 1/K$, the rollout takes the form starting from $\mathbf{x}^{(0)} = \mathbf{x}_0$.

$$t_k = \frac{k}{K}, \qquad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta t\, \mathbf{v}_\theta\Big(\mathbf{x}^{(k)}, t_k \mid \mathbf{c}\Big). \tag{3}$$

Unlike standard Flow Matching setups that sample $\mathbf{x}_0$ from an unconditional prior, our setting uses a query-induced initialization, while $\mathbf{x}_1$ is available only during training as the ground-truth query pose.

4 Method

Following Sec. 3, we instantiate conditional Flow Matching for Category-Agnostic Pose Estimation with three stages: heatmap-guided initialization, conditional flow matching, and probability-flow ODE rollout. Given a query image I_q, skeleton graph $\mathcal{G}$, and support point-description embeddings $\mathbf{T}_s = \{\mathbf{d}_i\}_{i=1}^N$, our goal is to predict the normalized query pose $\hat{\mathbf{x}} \in [0,1]^{N\times 2}$. The query image is encoded as $\mathbf{F}_q = \Phi_q(I_q)$, and the condition set used by the flow is $\mathbf{c} = \{\mathbf{F}_q, \mathbf{T}_s, \mathcal{G}\}$.

4.1 Heatmap-Guided Initialization

Instead of sampling $\mathbf{x}_0$ from an unconditional prior, we initialize the flow with a dedicated Heatmap-Guided Initialization module. It reuses the heatmap branch of **Riemannian Pose Head** and turns the input pair $(\mathbf{F}_q, \mathbf{T}_s)$ into a pose seed. The geometric anchor is the centered template $\mathbf{q}_0 = 0.5 \cdot \mathbf{1}_{N\times 2}$, where each keypoint is initialized at the image center in normalized coordinates.

To obtain a semantic prior, we query *Riemannian Pose Head* at $t = 0$ from $\mathbf{q}_0$ and extract a per-keypoint similarity map $\mathbf{H} \in \mathbb{R}^{N\times H_m \times W_m}$. Concretely, if $\widehat{\mathbf{f}}_{uv}$ denotes the ℓ_2-normalized query feature token at spatial location (u, v) and $\widehat{\mathbf{d}}_i$ denotes the normalized embedding of the i-th support point description, then the i-th heatmap channel is $H_{iuv} = \frac{\langle \widehat{\mathbf{d}}_i, \widehat{\mathbf{f}}_{uv} \rangle}{\tau_h}$, where $\widehat{\mathbf{d}}_i$ and $\widehat{\mathbf{f}}_{uv}$ are ℓ_2-normalized text and image embeddings, and τ_h is the heatmap similarity temperature.

We then convert $\mathbf{H}$ into continuous coordinates using soft-argmax:

$$\pi_{iuv} = \frac{\exp(H_{iuv}/\tau_s)}{\sum_{u',v'} \exp(H_{iu'v'}/\tau_s)}, \qquad \mathbf{x}_{0,i}^{\mathrm{hm}} = \sum_{u=1}^{H_m} \sum_{v=1}^{W_m} \pi_{iuv} \left[\frac{v-1}{W_m - 1}, \frac{u-1}{H_m - 1} \right]. \tag{4}$$

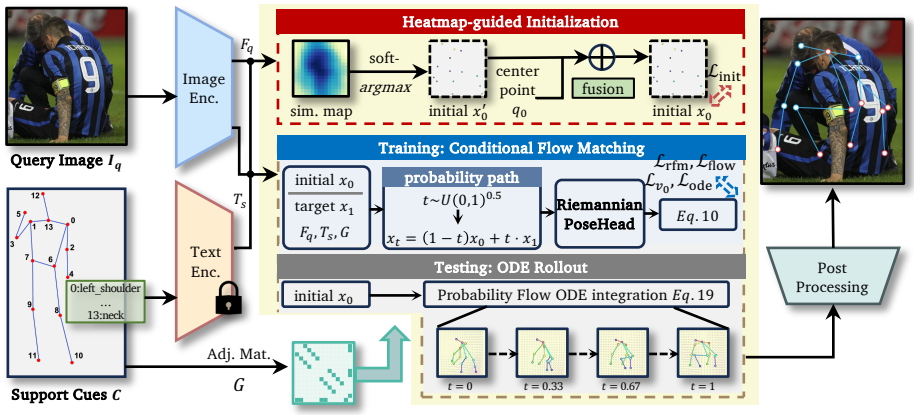

Fig. 2: Overview of the FlowCape. Image and text encoders extract query and support cues, a heatmap-guided initialization constructs the initial pose state, and a Riemannian Pose Head is trained with conditional flow matching. At test time, the pose is refined via ODE rollout to produce the final prediction.

where τ_s is the soft-argmax temperature. This yields the heatmap-based initialization $\mathbf{x}_0^{\text{hm}}$. The final initialization blends this semantic estimate with the centered anchor in Eq. (4):

$$\mathbf{x}_{0,i}^{\text{raw}} = \alpha_i \mathbf{x}_{0,i}^{\text{hm}} + (1-\alpha_i)\mathbf{q}_{0,i}. \tag{5}$$

This formulation covers pure heatmap initialization ($\alpha_i = 1$), fixed-ratio blending ($\alpha_i = 1-\beta$ with $\beta \in [0,1]$), and confidence-adaptive blending. In the confidence mode, the Heatmap-Guided Initialization module estimates the reliability of each keypoint from the heatmap peak and defines

$$p_i = \max_{u,v} H_{iuv}, \qquad \alpha_i = \sigma((\sigma(p_i) - \delta)\gamma)\,, \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function, and δ and γ denote the confidence threshold and slope, respectively.

From a data-flow perspective, the initialization stage consists of four steps: Image Encoder Φ_q extracts $\mathbf{F}_q$, Riemannian Pose Head produces similarity heatmaps, soft-argmax converts heatmaps into coordinates, and confidence-adaptive blending produces the final raw initialization $\mathbf{x}_0^{\text{raw}}$. During training, $\mathbf{x}_0^{\text{raw}}$ is supervised directly, but the state fed into the flow-matching objective is detached to avoid coupling the initialization gradients with the target velocity: $\mathbf{x}_0 = \text{sg}(\mathbf{x}_0^{\text{raw}})$, where $\text{sg}(\cdot)$ denotes stop-gradient. At inference time, we use $\mathbf{x}_0^{\text{raw}}$ directly as the initial state.

4.2 Conditional Flow Matching

Conditional velocity field. The second stage is implemented by the same *Riemannian Pose Head*, but now in vector-field prediction mode. Given an in-

termediate pose $\mathbf{x}_t$ and time $t \in [0,1]$, the head predicts a conditional velocity field $\mathbf{v}_\theta(\mathbf{x}_t, t \mid \mathbf{c})$. We first project the query features into a flattened query image feature $\mathbf{M} = \mathrm{Flatten}\left(W_f\mathbf{F}_q + \mathrm{PE}(W_f\mathbf{F}_q)\right)$, where W_f is a learnable projection and $\mathrm{PE}(\cdot)$ denotes the positional encoding. The i-th keypoint token is then initialized by combining support semantics and the current pose state,

$$\mathbf{z}_i^{(0)} = W_d\mathbf{d}_i + W_x\mathbf{x}_{t,i}, \tag{7}$$

where W_d and W_x are learnable projections. A sinusoidal embedding of t is injected inside the head before the stacked graph-flow blocks. We also compute a global text condition $\bar{\mathbf{d}} = \frac{1}{N}\sum_{i=1}^{N} W_d\mathbf{d}_i$, which modulates normalization layers via AdaLN and provides a global semantic context for all keypoints. The resulting velocity field is produced by the **Graph Flow Encoder**:

$$\mathbf{v}_\theta(\mathbf{x}_t, t \mid \mathbf{c}) = \mathrm{GFE}_\theta\left(\{\mathbf{z}_i^{(0)}\}_{i=1}^{N}, \mathbf{M}, \mathcal{G}, \bar{\mathbf{d}}, t\right). \tag{8}$$

Each Graph Flow block alternates self-attention over keypoint tokens, cross-attention to query image feature, graph smoothing over $\mathcal{G}$, and AdaLN modulation by $\bar{\mathbf{d}}$, allowing the model to couple local evidence with global structural priors. In this stage, current pose state $\mathbf{x}_t$ and support descriptions $\mathbf{T}_s$ form keypoint tokens, query features $\mathbf{F}_q$ form image feature, and GFE fuses them to predict the next transport direction.

Training objective. Let $\mathbf{x}_1 \in [0,1]^{N\times 2}$ denote the normalized ground-truth query pose. We follow the linear path in Sec. 3, but sample

$$t \sim \mathcal{U}(0,1)^{1/2}, \qquad \mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad \mathbf{u} = \mathbf{x}_1 - \mathbf{x}_0, \tag{9}$$

which emphasizes early-time dynamics near the initialization. Under the straight-path parameterization, the predicted velocity induces the terminal estimate

$$\hat{\mathbf{x}}_1 = \mathbf{x}_t + (1-t)\mathbf{v}_\theta(\mathbf{x}_t, t \mid \mathbf{c}). \tag{10}$$

For training, we use a validity mask $\mathbf{m} \in \{0,1\}^N$ to ignore unsupervised query keypoints. Let $\mathbf{r} = \mathbf{v}_\theta(\mathbf{x}_t, t \mid \mathbf{c}) - \mathbf{u}$ and $\mathbf{L}_\mathcal{G} = \mathbf{L}(\mathcal{G}, \mathbf{m})$ be the masked Laplacian metric induced by the skeleton graph. Our main supervision is a Hybrid Flow Matching objective,

$$\mathcal{L}_{\mathrm{rfm}} = \frac{1}{\sum_{i=1}^{N} m_i}\mathrm{tr}\left(\mathbf{r}^\top\mathbf{L}_\mathcal{G}\mathbf{r}\right), \qquad \mathcal{L}_{\mathrm{flow}} = \frac{1}{\sum_{i=1}^{N} m_i}\sum_{i=1}^{N} m_i \left\|\mathbf{r}_i\right\|_2^2, \tag{11}$$

where $\mathcal{L}_{\mathrm{rfm}}$ enforces structure-aware consistency under the skeleton metric and $\mathcal{L}_{\mathrm{flow}}$ preserves Euclidean regression accuracy.

We complement the velocity supervision with direct losses on the initialization and the recovered pose:

$$\mathcal{L}_{\mathrm{init}} = \frac{1}{\sum_{i=1}^{N} m_i}\sum_{i=1}^{N} m_i \left\|\mathbf{x}_{0,i}^{\mathrm{raw}} - \mathbf{x}_{1,i}\right\|_1, \tag{12}$$

To improve early-time accuracy and reduce distribution shift, we further impose $t = 0$ velocity loss and unrolled ODE consistency loss over K_{tr} steps:

$$\mathcal{L}_{\text{v0}} = \frac{1}{\sum_{i=1}^{N} m_i} \sum_{i=1}^{N} m_i \left\| \mathbf{v}_\theta(\mathbf{x}_0, 0 \mid \mathbf{c})_i - \mathbf{u}_i \right\|_2^2, \tag{13}$$

$$\mathbf{x}^{(0)} = \mathbf{x}_0, \qquad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \frac{1}{K_{\text{tr}}} \mathbf{v}_\theta\left(\mathbf{x}^{(k)}, \frac{k}{K_{\text{tr}}} \mid \mathbf{c}\right), \tag{14}$$

$$\mathcal{L}_{\text{ode}} = \frac{1}{\sum_{i=1}^{N} m_i} \sum_{i=1}^{N} m_i \left\| \mathbf{x}_i^{(K_{\text{tr}})} - \mathbf{x}_{1,i} \right\|_1. \tag{15}$$

The overall objective is then, and all λ terms are loss weights:

$$\mathcal{L} = \lambda_{\text{flow}} \mathcal{L}_{\text{flow}} + \lambda_{\text{rfm}} \mathcal{L}_{\text{rfm}} + \lambda_{\text{init}} \mathcal{L}_{\text{init}} + \lambda_{\text{v0}} \mathcal{L}_{\text{v0}} + \lambda_{\text{ode}} \mathcal{L}_{\text{ode}} \tag{16}$$

where $\mathcal{L}_{\text{hm}}$ is the auxiliary heatmap supervision from the shared head. When refinement layers are enabled, we keep the standard refinement loss $\mathcal{L}_{\text{ref}}$ from the prior refinement-based CAPE head and omit its unchanged formulation.

4.3 Inference via Probability Flow ODE Rollout

At test time, we initialize the state with the heatmap-guided pose $\mathbf{x}^{(0)} = \mathbf{x}_0$ and integrate the learned probability flow ODE with explicit Euler for K_{te} steps. At the k-th step, the shared Riemannian Pose Head is called as

$$t_k = \frac{k}{K_{\text{te}}}, \qquad \left(\mathbf{v}^{(k)}, \mathbf{H}^{(k)}\right) = \text{Head}_\theta\left(\mathbf{x}^{(k)}, t_k \mid \mathbf{c}\right). \tag{17}$$

We compute a heatmap potential $U(\mathbf{x}; \mathbf{H}^{(k)})$ by bilinearly sampling $\mathbf{H}^{(k)}$ at the current state and add its gradient to the velocity:

$$\mathbf{g}^{(k)} = \lambda_{\text{kg}} \nabla_{\mathbf{x}} U\left(\mathbf{x}^{(k)}; \mathbf{H}^{(k)}\right), \qquad \tilde{\mathbf{v}}^{(k)} = \mathbf{v}^{(k)} + \mathbf{g}^{(k)}. \tag{18}$$

The rollout update is

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \frac{1}{K_{\text{te}}} \tilde{\mathbf{v}}^{(k)}. \tag{19}$$

After K_{te} steps, we obtain the final normalized prediction $\hat{\mathbf{x}}$ and decode it to image coordinates by $\hat{\mathbf{p}}_i = \hat{\mathbf{x}}_i \odot [W_{\text{img}}, H_{\text{img}}]$. Thus, inference is formulated as an iterative probability-flow rollout, where the shared *Riemannian Pose Head* conditions on the current pose estimate, predicts the conditional velocity through its Graph Flow Encoder, and progressively advances the pose state until the final prediction is reached.

5 The Occ80 Benchmark

We introduce Occ80, an occlusion-centric multi-category benchmark for CAPE. Occ80 is constructed by selecting occlusion-heavy instances from MP-100 and integrating additional occluded samples from CrowdPose [14] and AnimalKingdom [22], resulting in 15,562 annotated instances across 80 categories. To better evaluate cross-category generalization under missing evidence, we further create 3 splits with mutually exclusive categories. More benchmark details are provided in the Supplementary Material.

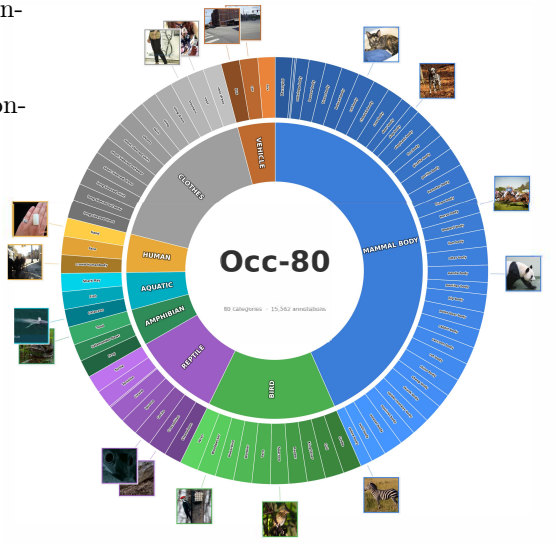

Fig. 3: Overview of the Occ80 benchmark.

6 Experiments

6.1 Settings

Implementation Details. We follow the CapeX [28] and use keypoint descriptors as the support cue. The architecture consists of an image encoder, a text encoder, and a Graph-aware Velocity Field Predictor. For fair comparison, we adopt Swin-V2 [19] as the visual encoder and a frozen gte-base-v1.5 [15] as the text encoder. The architecture is implemented in the MMPose framework [3] and trained with Adam for 200 epochs using a batch size of 16. The learning rate is initialized to 1×10^{-5} and decayed by a factor of 10 at epochs 160 and 180. All images are cropped and resized to 256×256 pixels. Training was conducted on a Linux server equipped with an NVIDIA A100 GPU. See more configurations in the Supplementary Material.

Datasets. We conducted our experiments on MP-100 dataset [33], the most prevalent and large-scale benchmark for CAPE, which covers 100 categories across 8 super-categories. MP-100 dataset contains keypoint numbers ranging from 8 to 68 across different classes. We follow the dataset splits in [33], *i.e.* all classes are split into non-overlapping train/val/test sets with the ratio of $70 : 10 : 20$, and there are five random splits to reduce the impact of randomness. In addition, to evaluate occlusion robustness in the most challenging settings, we perform extensive experiments on our proposed Occ80 benchmark. We use the same training protocol as in MP-100 for all 3 splits (S1-S3) of Occ80. Following the official implementations, we also reproduce CapeFormer [29], GraphCape [8], CapeX [28], GenCape [25], and EdgeCape [9] on Occ80.

Evaluation Metrics. We use the Probability of Correct Keypoint (PCK) as the evaluation metric. We follow the standard metric PCK@0.2 as the default

Table 1: Comparisons with the baselines on MP-100 dataset under 1-shot settings.

Method	Support	Split 1	Split 2	Split 3	Split 4	Split 5	Avg
POMNet [33]	Image	84.23	78.25	78.17	78.68	79.17	79.70
CapeFormer [29]	Image	89.45	84.88	83.59	83.53	85.09	85.31
ESCAPE [23]	Image	86.89	82.55	81.25	81.72	81.32	82.74
MetaPoint [1]	Image	90.43	85.59	84.52	84.34	85.96	86.17
GraphCape [8]	Image+Graph	91.19	87.81	85.68	85.87	85.61	87.23
SDPNet [27]	Image	91.54	86.72	85.49	85.77	87.26	87.36
PPM+CPT [24]	Image(+Text)	91.03	88.06	84.48	86.73	87.40	87.54
SCAPE [16]	Image	91.67	86.87	87.29	85.01	86.92	87.55
GenCape [25]	Image	92.05	88.69	86.89	85.88	87.02	88.09
EdgeCape [9]	Image+Graph	93.57	89.53	89.31	87.38	87.29	89.42
CapeX-T [28]	Text+Graph	**92.79**	**89.47**	84.95	87.25	**89.61**	**88.81**
FlowCape-T	Text+Graph	92.13	87.74	**85.18**	**87.79**	89.21	88.40
CapeX-S [28]	Text+Graph	95.62	90.94	88.95	89.43	92.57	91.50
FlowCape-S	Text+Graph	**96.19**	**90.97**	**89.19**	89.21	**92.71**	**91.65**

metric for performance reporting. We further evaluate our model using three standard keypoint localization metrics. AUC measures the area under the PCK curve and reflects overall accuracy across a range of distance thresholds. EPE computes the Euclidean distance between predicted and ground-truth keypoints. NME reports the mean localization error normalized by object scale. These metrics provide a comprehensive assessment of both absolute and scale-invariant localization performance. More results are in the Supplementary Material.

6.2 Benchmark Results

Quantitative Comparisons with Baseline Methods. Tab. 1 reports the evaluation metric scores for the MP-100 test set, considering the 1-shot scenario. As discussed in CapeX, we do not report 5-shot results because the support cue is a fixed keypoint text description, and using five shots would simply repeat the same text without providing additional information. We find that that our FlowCape achieves competitive performance with the strongest text-based baseline CapeX under the same backbone scale. In the Tiny setting, FlowCape-T obtains an average PCK of 88.40%, which is close to CapeX-T, while achieving better performance on Split 3 and Split 4. In the Small setting, FlowCape-S reaches an average score of 91.65%, slightly outperforming CapeX-S.

More importantly, Tab. 2 reports the PCK@0.2 results on the Occ80 dataset. Compared with prior CAPE methods, FlowCape consistently achieves the best performance. Specifically, with the SwinV2-Tiny backbone, FlowCape reaches an average score of 84.25%, outperforming the SOTA CapeX by +0.53%. The improvement becomes more pronounced with a larger backbone: SwinV2-Small,

Table 2: PCK@0.2 on the Occ80 dataset.

Backbone	Method	Src.	S1	S2	S3	Avg.
ResNet-R50	CapeFormer [29]	CVPR23	77.63	77.67	71.13	75.48
DINOv2-ViT-B/14	EdgeCape [9]	ICLR26	82.23	81.21	76.35	79.93
SwinV2-Tiny	GraphCape [8]	ECCV24	84.78	84.68	79.35	82.94
	CapeX [28]	ICLR25	85.68	85.79	79.69	83.72
	GenCape [25]	ICLR26	78.59	79.60	73.63	77.27
	FlowCape	-	**86.72**	**86.59**	**80.23**	**84.51**
SwinV2-Small	GraphCape [8]	ECCV24	84.97	85.23	79.92	83.37
	CapeX [28]	ICLR25	89.35	89.99	84.88	88.07
	GenCape [25]	ICLR26	87.61	84.35	80.66	84.21
	FlowCape	-	**91.56**	**90.49**	**85.32**	**89.12**

FlowCape achieves 89.12% Avg, surpassing CapeX by clear +1.05% margins. Notably, FlowCape attains the best performance across all 3 splits, demonstrating stable improvements across different category partitions. These comprehensive evaluations indicate that the proposed flow-based method effectively handles ambiguous and partially visible keypoints, allowing the model to maintain reliable structural predictions even when occlusions occur.

Qualitative Comparisons. Fig. 4 presents qualitative comparisons between CapeX [28] and our FlowCape under various challenging scenarios. As shown, severe occlusion and partial occlusion frequently cause ambiguous visual evidence for keypoints, leading to incorrect or structurally inconsistent predictions in CapeX. For instance, when keypoints are heavily occluded by objects or self-occlusion, *e.g.* limbs hidden behind the body or overlapping structures, CapeX often produces misplaced joints or distorted skeleton configurations. In contrast, FlowCape consistently recovers more coherent pose structures, even when several keypoints are partially invisible. This behavior is particularly evident in the leopard and human examples, where occluded limbs are still inferred with plausible spatial relationships. The error visualizations further highlight that FlowCape significantly reduces localization errors compared with CapeX. These results demonstrate that modeling pose generation through flow-based dynamics enables more ro-

Table 3: Ablations on components. 'H-init' represents Heatmap-guided Initialization. All experiments are conducted on Occ80 Split-1 with tiny backbone.

$\mathcal{L}_{\text{ODE}}$	H-Init	$\mathcal{L}_{\text{init}}$	$\mathcal{L}_{\text{rfm}}$	$\mathcal{L}_{\text{flow}}$	$\mathcal{L}_{v_0}$	**PCK**	Δ
						85.68	0
✓						72.81	↓13.31
✓	✓	✓				85.56	↓0.12
✓	✓	✓	✓			86.22	↑0.34
✓	✓	✓	✓	✓		86.58	↑0.90
✓	✓	✓			✓	86.41	↑0.73
✓	✓	✓	✓	✓	✓	86.72	↑1.04

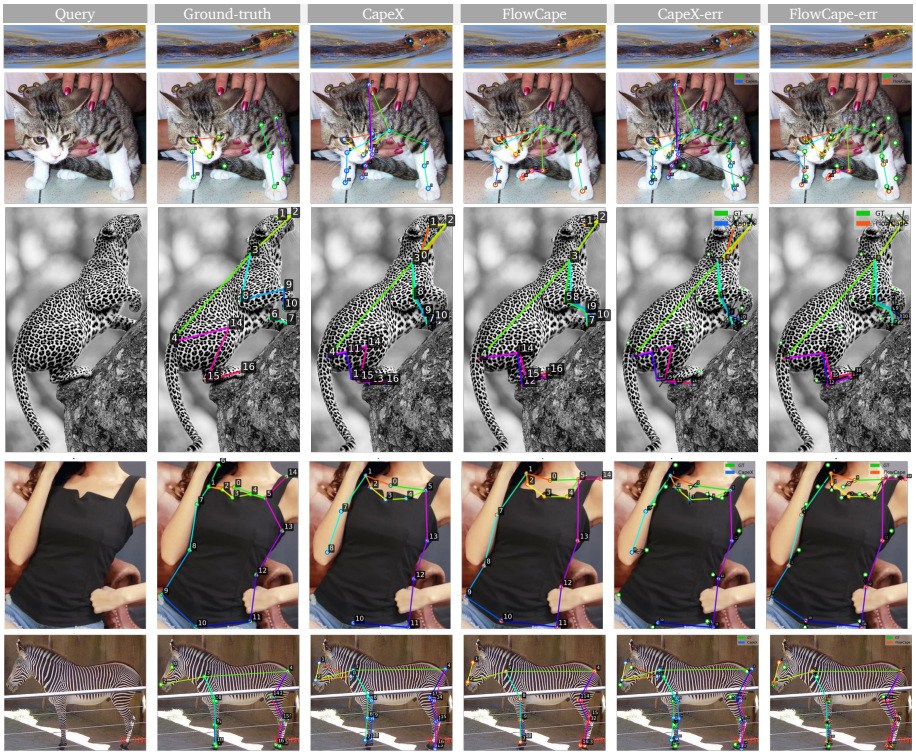

Fig. 4: Quantitative results of FlowCape on Occ80.From left to right, each row shows the input query image, ground truth, CapeX prediction, FlowCape prediction, the CapeX–GT error map, and the FlowCape–GT error map. The discrepancy between prediction and ground truth is indicated by double-ended black segments, which visualize the per-keypoint localization error.

bust pose reasoning under occlusion and partial visibility, leading to more stable and structurally consistent predictions.

6.3 Ablation Study

As the first attempt to apply flow matching to the CAPE task, our method warrants a careful analysis of each module, as well as the ODE integration and the initialization strategy. All ablation studies were conducted on the Occ80 Split-1 test set.

Different Components Analysis. We conducted experiments to evaluate the impact of various loss functions and the initial pose state x_0 strategy, including the initialization loss $\mathcal{L}_{\text{init}}$, the flow matching loss $\mathcal{L}_{\text{rfm}}$, the flow consistency loss $\mathcal{L}_{\text{flow}}$, the velocity regularization $\mathcal{L}_{v_0}$, and the ODE trajectory constraint $\mathcal{L}_{\text{ODE}}$. Tab. 3 summarizes the results. ODE-only means that we keep the continuous-time trajectory integration, but remove the auxiliary components.

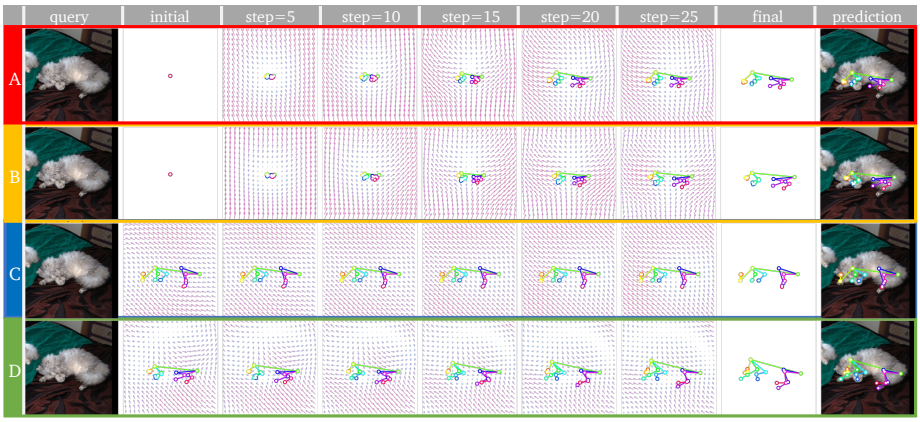

Fig. 5: Visualization of flow trajectories in different initialization strategy. A, B, C, and D correspond to the 20 steps query-centered, query-centered, heatmap-guided, and confidence-guided cases, respectively. The visualization is performed every five steps.

The large drop (-13.31%) is expected because ODE integration accumulates local errors along the trajectory. Under heavy occlusion, the posterior over poses is highly ambiguous and without flow-matching and consistency constraints the learned velocity field becomes weakly identifiable and prone to drift. Adding "H-Init" largely mitigates the drop by anchoring the trajectory near a plausible instance-specific pose, effectively shrinking the multi-modal occlusion posterior to a local neighborhood. This reduces ODE integration drift and yields cleaner supervision for learning a consistent velocity field, making the subsequent generative recovery well-conditioned. For flow dynamics, adding $\mathcal{L}_{\text{rfm}}$ and $\mathcal{L}_{\text{flow}}$ yields a substantial gain (+0.9%) because they provide direct supervision on the flow dynamics, aligning the learned velocity field with the target probability path and enforcing trajectory consistency under occlusion. The $\mathcal{L}_{v_0}$ term further improves performance by stabilizing the initial-time dynamics, reducing early-step drift that would otherwise be amplified during ODE integration. Overall, the results show that FlowCape benefits from the complementary effects of high-quality initialization and dynamics-consistent flow matching.

Different Initialization Mode Analysis. We study four strategies for constructing the initial pose state x_0 before flow-based refinement. (A) 20-step query-centered initializes x_0 at the query image center and runs the ODE for 20 steps; for visualization consistency in Fig. 5, the remaining steps are replicated to match the figure length. (B) Query-centered uses the same center-based initialization but runs the ODE for 30 steps, which is the default setting in all other experiments unless otherwise specified. (C) Heatmap-guided initializes x_0 by decoding the coarse heatmaps to obtain an initial state. (D) Confidence-guided

Table 4: Initialization strategy ablation on Occ80 Split-1.

Strategy	20S-Q.	Query	HM	Conf.
PCK	83.13	84.77	86.72	86.71

further weights the heatmap-derived coordinates by their response confidence, emphasizing high-confidence keypoints to form a more reliable initialization. As shown in Tab. 4, the performance consistently improves from left to right, suggesting that even a coarse-grained initialization is necessary to anchor the pose state and facilitate effective subsequent refinement. Fig. 5 corroborates this trend: center-based initialization is coarse and requires larger ODE trajectory corrections, while heatmap-/confidence-guided initialization injects query-grounded cues, starting closer to the correct mode and yielding shorter, more stable trajectories under occlusion. The near-tie between heatmap and confidence indicates diminishing returns once the initialization is already close, with confidence mainly helping suppress ambiguous peaks.

Different Hyper-parameter tuning. We further investigate the influence of several key hyper-parameters in the Riemannian Pose Head, including the pose head depth N_h, the hidden dimension D_h, and the weighting factors λ_{init} and λ_{traj}. As shown in Tab. 5, the performance improves as the pose head depth increases from 1 to 3 layers, reaching the best PCK of 86.72, while a deeper configuration (N_h=4) slightly degrades the performance, suggesting that that excessive depth may introduce redundant transformations. Similarly, increasing the hidden dimension from 128 to 256 improves the results, while further enlarging it to 384 yields only marginal gains at a noticeably higher computational cost. We therefore adopt $D_h = 256$ as the default. For the objective weights, the best performance is obtained when$\lambda_{\text{init}} = 10^{-1}$ and $\lambda_{\text{traj}} = 1$, suggesting that a balanced initialization constraint and trajectory supervision help stabilize the generative flow dynamics.

Table 5: Ablation studies on hyper-parameters, including the pose head depth N_h, the hidden dimension of the head D_h, and objective weights λ_{init} and λ_{traj} .

Hyper-parameters of Riemannian Pose Head				
Param. N_h (D_h=256)	1	2	3	4
PCK	83.51	86.54	**86.72**	84.96
Param. D_h (N_h=3)	128	192	256	384
PCK	85.61	85.89	86.72	**86.77**
Objective weighting factor				
Param. λ_{init} (λ_{traj}=1.0)	1.0	10^{-1}	10^{-2}	10^{-3}
PCK	86.58	**86.72**	86.19	86.23
Param. λ_{traj} (λ_{init}=10^{-1})	1.0	10^{-1}	10^{-2}	10^{-3}
PCK	**86.72**	86.28	85.67	84.79

7 Conclusion

We presented **FlowCape**, a conditional flow-matching framework for occlusion-resilient CAPE. By reformulating CAPE as structured pose transport, FlowCape improves prediction stability and global pose consistency under severe occlusion. We also introduce **Occ80**, an occlusion-focused benchmark for evaluating robustness and cross-category generalization. Experiments on MP-100 and Occ80 show that FlowCape consistently outperforms strong CAPE baselines, especially in challenging occluded scenarios.

Acknowledgements

This work was supported in part by the National Key R&D Program of China 2023YFC3806000 and 2023YFC3806002, in part by the National Natural Science Foundation of China under Grant U23A20382/62371342, in part by Shanghai Municipal Science and Technology Major Project No. 2021SHZDZX0100, in part by the Shanghai Science and Technology Innovation Action Plan Project 22511105300, in part by Shanghai Pujiang Program No. 25PJD128 and in part by the Fundamental Research Funds for the Central Universities. The authors would also like to thank the anonymous reviewers for their careful work and valuable suggestions.

References

1. Chen, J., Yan, J., Fang, Y., Niu, L.: Meta-point learning and refining for category-agnostic pose estimation. In: CVPR. pp. 23534–23543 (2024)
2. Chen, R.T., Lipman, Y.: Flow matching on general geometries. arXiv preprint arXiv:2302.03660 (2023)
3. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose` (2020)
4. Fei, H., Wu, S., Ji, W., Zhang, H., Chua, T.S.: Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In: CVPR. pp. 7641–7653 (2024)
5. Fu, Y., Yan, Q., Wang, L., Li, K., Liao, R.: Moflow: One-step flow matching for human trajectory forecasting via implicit maximum likelihood estimation based distillation. In: CVPR (2025)
6. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
7. Gui, M., Schusterbauer, J., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B.: Depthfm: Fast generative monocular depth estimation with flow matching. In: AAAI. pp. 3203–3211 (2025)
8. Hirschorn, O., Avidan, S.: A graph-based approach for category-agnostic pose estimation. In: ECCV. pp. 469–485. Springer (2024)
9. Hirschorn, O., Avidan, S.: Edgecape: Edge weight prediction for category-agnostic pose estimation. In: ICLR (2026)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. vol. 33, pp. 6840–6851 (2020)
11. Holmquist, K., Wandt, B.: Diffpose: Multi-hypothesis human pose estimation using diffusion models. In: ICCV. pp. 15977–15987 (2023)
12. Kim, J., Chung, H., Kim, B.H.: Capellm: Support-free category-agnostic pose estimation with multimodal large language models. ICCV (2025)
13. Le, C., Melnyk, P., Wandt, B., Wadenbäck, M.: Flow matching for probabilistic monocular 3d human pose estimation. arXiv preprint arXiv:2601.16763 (2026)
14. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: CVPR. pp. 10863–10872 (2019)
15. Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M.: Towards general text embeddings with multi-stage contrastive learning (2023)

16. Liang, Y., Ye, Z., Liu, W., Lu, H.: Scape: A simple and strong category-agnostic pose estimator. In: ECCV. pp. 478–494. Springer (2024)
17. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling (2023)
18. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow (2022)
19. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: CVPR. pp. 12009–12019 (2022)
20. Liu, Z., Zhou, S., Liu, M., Deng, H., Zhu, H.: Prismprune: Decoupling saliency and diversity in attention for efficient visual token pruning in vlms. In: CVPR Findings. pp. 6174–6184 (2026)
21. Mueller, A.: Modern robotics: Mechanics, planning, and control [bookshelf]. IEEE Control Systems Magazine **39**(6), 100–102 (2019)
22. Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J.: Animal kingdom: A large and diverse dataset for animal behavior understanding. In: CVPR. pp. 19023–19034 (2022)
23. Nguyen, K.D., Li, C., Lee, G.H.: Escape: Encoding super-keypoints for category-agnostic pose estimation. In: CVPR. pp. 23491–23500 (2024)
24. Peng, D., Zhang, Z., Hu, P., Ke, Q., Yau, D.K., Liu, J.: Harnessing text-to-image diffusion models for category-agnostic pose estimation. In: ECCV. pp. 342–360 (2024)
25. Rao, J., Wang, Y., Zhao, S.: Gencape: Structure-inductive generative modeling for category-agnostic pose estimation. In: ICLR (2026)
26. Rao, J., Zhao, B.N., Wang, Y.: Probabilistic prompt distribution learning for animal pose estimation. In: CVPR. pp. 29438–29447 (2025)
27. Ren, P., Gao, Y., Sun, H., Qi, Q., Wang, J., Liao, J.: Dynamic support information mining for category-agnostic pose estimation. In: CVPR. pp. 1921–1930 (2024)
28. Rusanovsky, M., Hirschorn, O., Avidan, S.: Capex: Category-agnostic pose estimation from textual point explanation. In: ICLR (2025)
29. Shi, M., Huang, Z., Ma, X., Hu, X., Cao, Z.: Matching is not enough: A two-stage framework for category-agnostic pose estimation. In: CVPR. pp. 7308–7317 (2023)
30. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2021)
31. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703 (2019)
32. Wang, T., Yu, X., Mathis, M.W.: Fmpose3d: monocular 3d pose estimation via flow matching. In: CVPR (2026)
33. Xu, L., Jin, S., Zeng, W., Liu, W., Qian, C., Ouyang, W., Luo, P., Wang, X.: Pose for everything: Towards category-agnostic pose estimation. In: ECCV. pp. 398–416. Springer (2022)
34. Xu, T., Rao, J., Song, X., Feng, Z., Wu, X.J.: Learning structure-supporting dependencies via keypoint interactive transformer for general mammal pose estimation. IJCV pp. 1–19 (2025)
35. Yang, J., Zeng, A., Zhang, R., Zhang, L.: X-pose: Detecting any keypoints. In: ECCV. pp. 249–268 (2024)
36. Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: Hrformer: High-resolution vision transformer for dense predict. In: NeurIPS. vol. 34, pp. 7281–7293 (2021)