We thank the reviewers' insightful comments and address the raised issues below. We will revise our paper based on these comments.

[UPbE]C1-Experiment Issues. (Q5) **(i)** For synthetic datasets, AP-10k is randomly & evenly split into two subsets, and the joints division is in Fig A. The details are introduced in Lines 737-744. For real dataset combination, we first merge multiple datasets and randomly select 10% as our labeled data. The random seed for all selection is 2. **(ii)** Despite training with scarce incomplete annotations, there is no risk of overfitting due to the use of unlabeled data. The obtained results show consistent improvement of ours on both datasets.

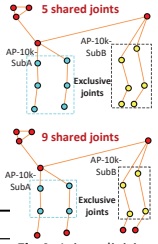

Fig A: Joints division

mAP/PCK@0.05	Model 1:Ls	Model 2:Ls+Lu	Ours
AP-10k	52.1/66.3	56.2/69.3	57.2/70.1
AnimalPose	57.3/65.5	61.7/69.6	62.4/69.6

(iii) We design a study to evaluate the effect of different dataset combinations. We follow the setting in Table 4. The results show that ours consistently outperforms SL by a large margin under various combination ratios, and a larger ratio leads to higher accuracy.

mAP/PCK@0.05	5% comb. ratio	10% comb. ratio	20% comb. ratio
SL	42.2/60.0	52.2/67.6	60.3/72.4
FreeNet(Ours)	48.6/65.3	57.26/71.36	63.4/75.4

C2-Enhance our framework. (Q4) We add details such as the learning targets, math formulas, and the selection criteria in Fig B.

C3-Explain $\mathcal{L}_u$, $\mathcal{L}_s$, $\mathcal{L}_f$. (Q2) Eq1: $\mathcal{L}_u$ is similar to the classic unsupervised loss, which measures the difference between pseudo-GT and the adaptation network's predictions for unlabeled data. The key difference lies in using joints k in $\hat{\mathbb{U}}$ that meet the body part-aware sampling criteria for loss calculation, denoted as $\sum_{k=1}^{N_J}\{\hat{H}_u^k \in \hat{\mathbb{U}}\}_1 \mathcal{L}_u^k$. For enhanced training stability, it is further scaled by the number of selected joints. Eq2: $\mathcal{L}_s$ is the supervised loss applies to the merging of M non-standard datasets. Eq3: $\mathcal{L}_f$ measures the difference between pseudo-GT and the base network's predictions for unlabeled data, multiplied by the feedback factor f. First, the value of f (Eq4) is a dot product of two gradients terms, representing the cognitive differences of the adaptation network on labeled and unlabeled data. Second, pseudo joints k in $\hat{\mathbb{S}}_{feedback}$ are used for loss calculation, $\sum_{k=1}^{N_J}\left\{\hat{H}_u^k \in \hat{\mathbb{S}}_{feedback}\right\}_1 \mathcal{L}_f^k$, to better predict unannotated joints. Third, the loss is also scaled to ensure training stability. (**cf.Zt6j.C2**)

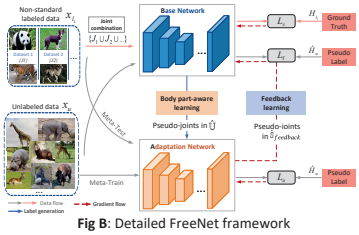

Fig B: Detailed FreeNet framework

C4-Clarification. (Q3) We specify pseudo-label for pose estimation as pseudo-joints, denoted as $\hat{H}_u$, where $\hat{H}_u^k$ represents the kth joint. Unlike GT labels, pseudo labels or joints change dynamically during training. The way to obtain them is detailed in Lines 449-453.

C5-Improve fluency and logical clarity. (Q1) Thanks for your suggestions. We will thoroughly proofread and polish our paper to improve its presentation. To enhance logical clarity, we will clearly define key terms such as "pseudo-joints", update the framework Figure 3(a) with Fig B, and better explain the math formulas.

[niDm]C1-Will different categories of keypoints cause a long-tail effect? (Q1) Pose variations and occlusions obscure some keypoints, leading to an inherent long-tail distribution that is further aggravated by the inclusion of unannotated joints. *The long-tail effect persists regardless of applying different categories of keypoints.* Three experiment results show that FreeNet effectively mitigates this issue. First, it significantly improves the accuracy of tail keypoints in Fig C. Second, it reduces the accuracy gap between 5/9 shared joints and unannotated joints from 10.4/8.7 to 8.46/7.5 (see Table 3). Third, it improves the std accuracy of three body parts by 0.6 (see Table 5).

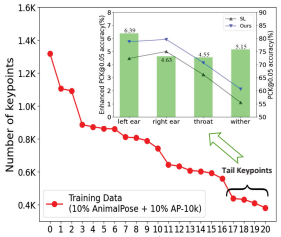

Fig C: Mitigating long-tail effect

C2-Gradient propagation for missing points. (Q2) Missing points arise from two issues: regular occlusions and lack of standard annotation definitions across datasets. Unlike occluded points, the second type of missing points are not completely unannotated; instead, they are marked as 0 when unannotated. Only those annotated points are forward-passed for $\mathcal{L}_s$, and then gradients are propagated to update the base network. **0-valued missing points, akin to occluded points, are excluded from loss calculation and gradients update** since their GT labels are not available. The accuracy of unannotated missing points is further improved through body part-aware learning and feedback learning in FreeNet.

C3-Clarification. (Q3) **(i)** Yes, the learned joints match non-standard labeled data properties (see Table 1). **(ii)** From top to bottom, these three keypoints represent the wither, neck, and throat.

[Zt6j]C1-Detailed method steps. (Q1) Figure 3 (or Fig B) presents the training process of FreeNet. In the inference, the adaptation network is used for pose estimation. The detailed steps can be found in Lines 341-385, and we will clearly revise it as suggested.

C2-About feedback learning. (Q2)**(i)** Its update process in T steps: 1) Sample a batch of x_l and x_u from given dataset D, 2) Establish $\hat{\mathbb{S}}_{feedback}$ using the threshold $\alpha_{feedback}$, 3) Calculate feedback factor f based on the difference between "new" and "old" adaptation networks on x_l and x_u, respectively. 4) Use x_u in $\hat{\mathbb{S}}_{feedback}$ to update the base network $\theta_{\mathcal{B}}$ based on $\mathcal{L}_f$. **(ii)** The update is simple to execute, as only the base network is updated directly. Although the f calculation involves the adaptation network in two steps, it does not require gradients update. **(iii)** Sorry, we recognize that "distillation" may be confusing as our two models are of the same size. We believe "refinement" is more appropriate to show that the predictions of unannotated joints are refined via feedback learning.

C3-Discuss more limitations. (Q3) When tackling unseen extremely diverse or rare species, FreeNet generally performs well in predicting shared joints because these joints appear in many species, facilitating effective feature transfer. If there are many species with exclusive joints available, FreeNet is likely to predict dense joints for unseen species accurately. However, if only a few species have such joints, especially when these species have little similarity to the unseen ones, FreeNet may have difficulty generalizing to exclusive joints. Dense joint prediction for unseen species is valuable but more challenging than general animal pose estimation or dense joint prediction for known species. We will add it to our revision.

C4-Other issues. Q4: We will clearly revise the Figure 6 as suggested. Q5: According to Lines 721-724, all compared methods in Table 2 have the same model complexity (28.5M Param) because they all utilize HRNet-w32 as the backbone network.