Incremental Object Keypoint Learning

Supplementary Material

Contents

1. Guideline of Supplementary

In this supplementary, in Sec. 2, we first provide more discussion of the non-co-occurrence (NCO) problem among Incremental Object Detection (IOD), Incremental Semantic Segmentation (ISS), and Incremental Object Keypoint Learning (IKL) to highlight the novel contribution to resolving the NCO issue in IKL.

in Sec. 3, we first provide further discussions on the different purposes of our two distillation losses designed for the Stage-II training in our proposed KAMP method, i.e., our Knowledge Association loss ℓ_{KA} created by the KA-Net and our Keypoint Spatial-oriented Distillation loss ℓ_{KSD} created by the old model m_{t-1}.

In Sec. 3.1, we empirically compare the performance of the KA-Net and the old model m_{t-1} on predicting the old keypoints selected in Stage-I and show that they are **not** similar, as the anatomical prior captured in our KA-Net can further improve the estimations of related old keypoints. This further verifies that the loss ℓ_{KA} by KA-Net and the loss ℓ_{KSD} by the old model may perform different kinds of knowledge distillation in Stage-II.

As mentioned in our **main paper Sec. 4.2**, in Sec. 3.2 and 3.3, we provide the ablation study on the Head-2023 dataset and also test our method on other network backbone. In Sec. 3.4, we provide further ablation study that trains the KA-Net with the ground truth label of the selected old keypoint instead of using the old model m_{t-1} to provide the pseudo-label, where their performances are **almost the same**, demonstrating that it is **not** critical to use completely accurate labels to train KA-Net to achieve great results for our method in IKL.

In Sec. 3.5, we also provide further ablation study on only using the loss ℓ_{KA} without the loss ℓ_{KSD} in Stage-II training of our method, where we show that it is necessary

to use both the loss ℓ_{KA} and ℓ_{KSD} simultaneously given their different functionalities and complementary property.

Then, we discuss the limitations of our proposed method in Sec. 4, as mentioned in our **main paper Sec. 4.3**. In Sec. 5, we discuss the extreme case when only one new keypoint is considered for constructing the KA-Net in Sec. 5.1 as mentioned in our **main paper Sec. 3.2.1**, and more details of the KA-Net. We further provide concrete examples of constructing the auxiliary task in Sec. 5.3, and provide discussion and ablation study of different Softmax alternatives in Sec. 5.4 and 5.5 respectively.

As mentioned in our **main paper Sec. 4**, starting from Sec. 6, we provide more dataset statistics and experimental details in Sec. 6.2 and include per-step performance and the standard deviation of each dataset in our main paper's Table 1 in Sec. 6.4. In Sec. 6.1, we provide more details of our evaluation metric as mentioned in our **main paper Sec. 4**. In Sec. 6.3, we provide more details of our adaptation for the low-shot regime as mentioned in our **main paper Sec. 4.3**. The details of the keypoint group and analysis of the α are included in Sec. 6.6 and Sec. 6.5 respectively, as mentioned in our **main paper Sec. 4**. We also report more results over a balanced experimental protocol in Sec. 6.7, and another challenging experimental protocol when old keypoints are all from the human upper body and new keypoints are all from the human lower body in Split MPII in Sec. 6.8. Then we provide more experimental details about the comparison experiments between IKL and other alternative settings (i.e., CC2D [38], EGT [39], UKL [13], and CAPE [29]) in Sec. 6.9 as mention in **main paper Sec. 4.3**. Finally, we present the per-keypoint transfer metric of Split ATRW after three incremental steps in Sec. 6.10 as mentioned in our **main paper Sec. 4.1**, and more visualization results in Sec. 6.11.

2. More Discussion of the non-co-occurrence (NCO) problem among Incremental Object Detection (IOD), Incremental Semantic Segmentation (ISS) and Incremental Object Keypoint Learning (IKL)

While the non-co-occurrence (NCO) problem exists in IOD, ISS, and IKL, the nature and implications of NCO are fundamentally different: (1) In IOD/ISS, NCO affects discrimination between distinct object categories (e.g., cat & dog). The challenge is primarily about maintaining class boundaries when old classes are not labeled in new data. (2) In IKL, NCO manifests within the same object category, where new and old keypoints have inherent anatomical and physical relations. The challenge is not class discrimination, but capturing these intrinsic relationships to improve keypoint estimation.

This unique context of NCO in IKL motivated our novel technical solution: KAMP explicitly models and leverages the relationships between old and new keypoints, improving the estimation of both beyond mere anti-forgetting. This approach is specifically designed for the keypoint estimation context and differs fundamentally from IOD/ISS solutions.

3. Different Proposes of Knowledge Distillation between ℓ_{KA} and ℓ_{KSD} in the Stage-II.

As stated in our main paper Sec. 3.2.1, in Stage-II training, the Knowledge Association loss ℓ_{KA} is created by the frozen KA-Net, which is learned in Stage-I to capture the implicit anatomical and physical prior between the related old and new keypoint. We leverage ℓ_{KA} in Stage-II to distill the **keypoint association knowledge** to further improve **the selected old keypoints** in $\mathcal{K}_{KA}$ defined in our main paper Sec. 3.2.1.

While for the loss ℓ_{KSD}, as stated in our main paper Sec. 3.2.2, since the loss ℓ_{KA} only applies to **the selected old keypoints** to distill their keypoint association knowledge instead of mitigating the forgetting of all the keypoint, thus the loss ℓ_{KSD} is used to distill **the old model's knowledge** during Stage-II training to consolidate the knowledge for **all the keypoints** to avoid catastrophic forgetting.

Therefore, in Stage-II training, for **the selected old keypoint** (orange in Fig. 1), they are supervised by both the loss ℓ_{KA} and ℓ_{KSD}, where the loss ℓ_{KA} is for distilling their new knowledge of the keypoint association with the related new keypoints, and the loss ℓ_{KSD} is for distilling their previous knowledge from the old model m_{t-1}. For other old keypoints, they are all supervised by the loss ℓ_{KSD} for knowledge consolidation.

One may be concerned about two factors: (1) As both the KA-Net and old model can predict **the selected old keypoints**, it would be interesting to see whether the KA-Net may be similar to the old model m_{t-1} on predicting **the selected old keypoints** such that the KA-Net may not provide a different distillation effect for the **the selected old keypoints** compared to the old model during Stage-II. (2) As the KA-Net is trained by the pseudo-label provided by the old model m_{t-1} in Stage-I since we will not label the previously-learned old keypoints during IKL, then given that the prediction from the old model is not as perfect as the ground-truth label, it would be also interesting to see whether the performance of our propose KAMP method may be influenced by the training of the KA-Net.

To study these two concerns, we first empirically explore the first concern in Sec. 3.1 by comparing the performance of **the selected old keypoints** based on the KA-Net and the old model m_{t-1}. We show that the KA-Net is **not** similar to the old model, and the KA-Net can achieve much better performance on **the selected old keypoints** than the old model. Then in Sec. 3.4, we train the KA-Net with the ground-truth

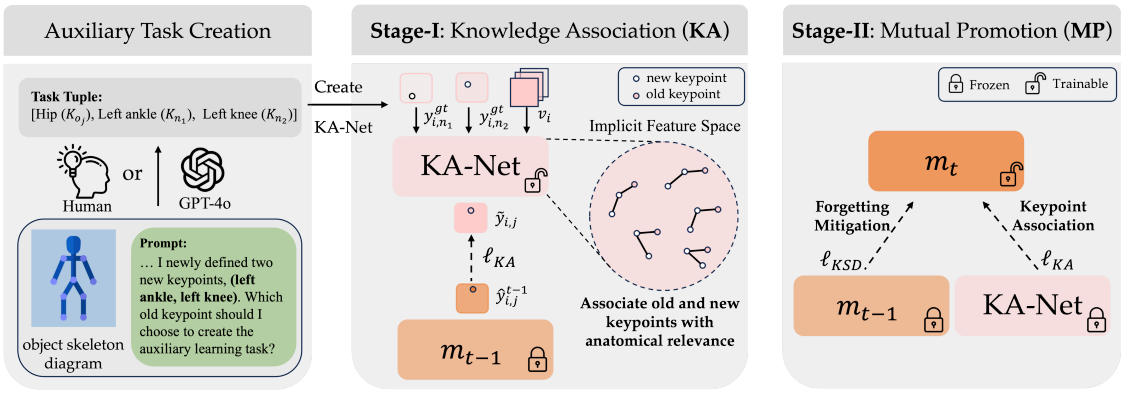

Figure 1. Overview of KAMP using the human body for illustration. In Stage-I, we learn an auxiliary KA-Net to associate the related old and new keypoints based on their local anatomical constraint. In Stage-II, we jointly leverage the old model and the KA-Net as an auxiliary teacher to consolidate all old keypoints' prediction and also learn the new keypoints simultaneously to achieve mutual promotions.

labels of the old keypoints to see whether it is critical to use very accurate and perfect supervision for training the KA-Net. Our empirical results demonstrate that there is **almost no difference** between the performance of our KAMP when the KA-Net is either supervised by the pseudo-label from the old model m_{t-1} or by the ground-truth label, which implies that it is **not** critical to use completely accurate labels to train the KA-Net to achieve great results.

3.1. Difference between the Performance of the KA-Net and the Old Model on Predicting the Old Keypoints Selected in Stage-I.

As mentioned in Sec. 3, here we explore the performance difference between the KA-Net and the old model m_{t-1} on predicting **the selected old keypoints** by using the 5-Step Split MPII protocol. As shown in the Tab. 1, our KA-Net is **not** similar to the old model on predicting the selected old keypoint, and our KA-Net are all better than the old model on predicting the old keypoints selected in each incremental step. This is because, as stated in our main paper Sec. 3.2, the selected old keypoint predicted by the KA-Net is conditioned on the newly-defined keypoints that are not defined when training the old model. Thus, compared to the prediction from the old model, the prediction of the selected old keypoint from the KA-Net is supplemented with the new anatomical knowledge between the old and new keypoint. Such a physical prior can support the old keypoint prediction by constructing a local constraint to improve the prediction robustness of the old one, and thus, the prediction from the KA-Net for the selected old keypoint is better than the corresponding prediction from the old model.

Therefore, given the difference between the KA-Net and the old model, the KA-Net can further serve as an auxiliary teacher that is different from the old model in Stage-II training to further improve the performance of our KAMP method, which has been verified in our ablation study in our main paper's Sec. 4.2 and Tab. 2.

SOK in each step	Step-1	Step-2	Step-3	Step-4
Old Model m_{t-1}	64.39	68.27	77.32	95.70
KA-Net	**67.36**	**72.78**	**83.90**	**96.32**

Table 1. Comparison between the old model m_{t-1} and our KA-Net on predicting **the Select Old Keypoint** (SOK) in each incremental step in 5-Step Split MPII.

3.2. Ablation study on Head-2023 (main paper Sec. 4.2)

As mentioned in our main paper Sec. 4.2, here we replicate the ablation study of Tab. 2 in our main paper on Split Head-2023. As shown in Tab. 2, we can achieve the same conclusion stated in our main paper's Sec. 4.2.

Method	A-MRE$_4$ ↓	AT$_4$	MT$_4$
LWF [24]	4.31	-1.26	0.57
KAMP (only ℓ_{KSD})	3.29	-0.12	0.63
KAMP (Random KA-Net)	2.93	0.08	0.73
KAMP (Ours)	**2.32**	**0.41**	**0.84**

Table 2. Ablation Study on Split Head-2023

3.3. Ablation study on different backbone other than HR-Net (main paper Sec. 4.2)

As stated in our main paper Sec. 3, our KAMP design is general, versatile, and usable with various backbones [3, 14, 15, 23, 25, 35, 36]. Here, we leverage one of the state-of-the-art methods, i.e., Residual Steps Network (RSN) [3], as our backbone, and we can achieve 81.45% AAA$_4$ on Split MPII, which further outperform 79.93% obtained by using HRNet backbone in our Table 1 in the main paper. This verifies the generality of our proposed KAMP method.

3.4. Train KA-Net with Ground-truth Labels.

As mentioned in Sec. 3, here we further explore whether it is critical to use completely accurate supervision to train the KA-Net to achieve a great result for our KAMP method. As shown in Tab. 3, we can observe that even if we use the ground-truth labels of the selected old keypoint to train the KA-Net, the overall average performance (AAA_4) of 5-Step Split MPII (79.97%) is almost the same as Ours (79.93%) that uses the pseudo-label provided by the old model m_{t-1}. This shows that it is **not** so critical to use completely accurate labels to train the KA-Net to achieve great results.

	5-Step Split MPII		
Method	AAA_4	AT_4	MT_4
Ours (KA-Net trained by GT)	**79.97**	1.70	**5.63**
Ours	79.93	**1.80**	4.23

Table 3. Training the KA-Net by using the pseudo-label provided by the old model m_{t-1}, i.e., Ours, and by using the corresponding ground truth (GT) labels, i.e., Ours (KA-Net trained by GT).

Method	AAA_4	AT_4	MT_4
LWF [24]	75.75	-3.86	0.41
KAMP (only ℓ_{KA})	54.46	-12.74	-3.45
KAMP (only ℓ_{KSD})	76.93	-2.24	0.65
KAMP (Ours)	**79.93**	**1.80**	**4.23**

Table 4. More Ablation Study on 5 Step Split MPII.

3.5. The Necessity of Using ℓ_{KA} and ℓ_{KSD} Simultaneously.

As we mentioned in our main paper Sec. 3.2.2, we emphasize that only the loss ℓ_{KA} is not enough to mitigate the forgetting of **all the old keypoints**, since the loss ℓ_{KA} is only applied to **the selected old keypoint** and its functionality is only to distill the keypoint association knowledge to improve the predictions of selected old keypoints. Thus the usage of the loss ℓ_{KSD} is necessary as its functionality is to consolidate the knowledge of all the old keypoints based on the old model m_{t-1} to mitigate the forgetting problem. To verify this claim, we extend the ablation study in our main paper Sec. 4.2 and Tab. 2, where we add one more experiment, i.e., only using the ℓ_{KA} without the ℓ_{KSD} in Stage-II training. The results are shown in Tab. 4, and we can observe that with only ℓ_{KA} in Stage-II training, the old keypoints are catastrophically forgotten on 5-Step Split MPII. Only when we use both the loss ℓ_{KA} and ℓ_{KSD} simultaneously, i.e., Ours, can we achieve the best result. This demonstrates that the different functionalities of the loss ℓ_{KA} and ℓ_{KSD} make it necessary to use them simultaneously such that we can effectively mitigate the catastrophic forgetting of the old ones and then further improve them. By comparing the alternative of only using the loss ℓ_{KSD} in Stage-II and Ours, we can see that the loss ℓ_{KA} is complementary to the loss ℓ_{KSD} as adding the loss ℓ_{KA} to the loss ℓ_{KSD} can help us further achieve larger improvement on both the average performance (i.e., AAA_4) and the old keypoints (i.e., AT_4 and MT_4). This further implies the complementary property of the loss ℓ_{KSD} and ℓ_{KA}.

4. Discussions of the Limitations. (main paper Sec. 4.3)

As mentioned in our main paper's Sec. 4.3, since we do not have the ground-truth labels for the old keypoints, thus we leverage the old model's prediction of the old keypoints as the pseudo-label to supervise the KA-Net. However, there would be a concern that the old model's prediction may not be accurate enough to provide supervision as good as the ground-truth label for the old keypoints.

Our discussions towards this concern are two-fold: (1) in this paper, we actually do not need the KA-Net to be perfect enough to be an auxiliary teacher for the old keypoints. Instead, we hypothesize that even a not strong enough teacher like KA-Net could already benefit the IKL since the KA-Net has different functionality, i.e., it is used to implicitly distill the knowledge of keypoint association into the new model during the IKL. Our empirical study in both the main paper and our supplementary all verified our hypothesis since all the experiments were conducted without concern about whether the KA-Net is strong enough to predict the old keypoints. The empirical result in Sec. 3.4 and Tab. 3 further supports that the completely accurate supervision for training the KA-Net is **not** critical to achieving great results for our method in IKL. (2) In the future, we can further explore leveraging the technique of uncertainty estimation to filter out those uncertain predictions from the old model when training the KA-Net, such that we can prevent the potential low-quality old keypoints' predictions from unstabilizing the training of KA-Net.

Lastly, regarding our adaptation to the low-shot regime, we employ the same training strategy as outlined in [38] to train an auxiliary model in a self-supervised manner. This model is then used to pseudo-label new keypoints when only a few annotations are available in the new data. Consequently, as demonstrated in Tables 3 and 4 of our main paper, the quality of our pseudo-labeling may be constrained by the limitations of [38] in extreme 1-shot and/or 5-shot scenarios. We anticipate that these limitations could be overcome with future developments in self-supervised learning. Furthermore, it's important to highlight that our adaptation strategy is primarily introduced to demonstrate the feasibility of our KAMP method in extremely low-shot conditions and to provide a comparative analysis with other low-shot methods. However, in practical applications where

high accuracy in keypoint detection is crucial, such as in medical analysis, we generally prefer algorithms that can scale performance with an increase in available labels. In this regard, our KAMP method is advantageous over other alternatives. It not only performs better in low-shot scenarios but also scales more effectively with additional labeled data. For example, labeling 10 to 50 images, which is a manageable task even in medical contexts, can significantly enhance the training of a reliable keypoint detector. As shown in Tables 3 and 4 of our main paper, our KAMP method uniquely scales up with more labels, making it a more favorable option for real-world applications compared to other alternatives.

5. More Analysis of the Proposed Method.

5.1. Extreme Case When Only Considering One New keypoint for Constructing the KA-Net. (main paper Sec. 3.2.1)

As mentioned in our main paper Sec. 3.2.1, the auxiliary task construction for the KA-Net can also be created like this: $P(K_j^{\text{old}}) = F(P(K_1^{\text{new}}), P(K_i^{\text{old}}))$, where when we only have one new keypoint K_1^{new} that is related to the old keypoint K_j^{old}, we can consider other old keypoint K_i^{old} that it is also related to K_j^{old} and K_1^{new}. This shows the generality of our proposed method in that it is still feasible when the extreme case occurs, e.g., only one new keypoint is introduced. To empirically verify the feasibility, we provide a case study using Split MPII and only introduce one new keypoint in each incremental step, as shown in Tab. 5. We can observe that our method can still achieve a positive average transfer and positive maximal transfer under such an extreme scenario, where we still outperform the competitive baseline, i.e., LWF [24], with a large margin.

	5-Step Split MPII		
Method	AAA_4	AT_4	MT_4
LWF [24]	75.75	-3.86	0.41
KAMP (Ours)	**79.14**	**1.41**	**4.74**

Table 5. Results when only one new keypoint is considered for constructing the KA-Net.

5.2. More Details of the KA-Net

As described in our main paper's Sec. 3.2.1, we extract the spatial-oriented features for each new keypoints that are the input of the KA-Net. As we want the KA-Net to capture the keypoint association between the related old and new keypoint, thus the KA-Net needs to have the capability of modeling the spatial correlation between those keypoints. Therefore, we use a large convolution kernel to capture the long-range correlation between the old and new keypoints' visual features: for the first two convolutional layers in KA-Net, their kernel size is 15×15, and the padding size is 7. The last convolution layer is for generating the heatmap prediction of the old keypoint, and hence its kernel size is 1×1 and padding size is 0.

5.3. Concrete Examples of Task Constructions.

In our main paper Sec. 3 and as shown in Fig. 1, when we consider a category for keypoint estimation, it is feasible to create an auxiliary prediction task based on a general object anatomy diagram. These diagrams are readily available on the internet, and a human can interpret them by understanding the semantic meanings of both old and new keypoints. In some medical imaging applications, a doctor may need to clarify the definition of keypoints, but this is a **one-time** task and is significantly faster than having a doctor label every image.

Utilizing a general object anatomy diagram for keypoint association eliminates the need to label training images to learn which keypoints should be associated, thus saving substantial annotation and training costs. This approach also provides an interpretable and flexible method for humans to apply physical knowledge of an object category to facilitate incremental learning for the first time. We compare this method with an alternative that involves feeding new data to the old model to obtain pseudo-locations of all old keypoints, and then measuring the relative distances between new and old keypoints to create the auxiliary task. Our method not only reduces the time required from 5.42 minutes to 0.594 seconds but also improves the AAA_4 from 78.82% to **79.93%** for the Split MPII dataset. This improvement is due to the fact that the pseudo-locations predicted by the alternative method may be incorrect, leading to improper creation of the auxiliary task.

To provide concrete examples, Fig. 2 (a), (b), and (c) show that an object anatomy diagram can be easily found online. Since we use this diagram only to identify the proximity of relative locations between old and new keypoints, it can be quite general. Based on this diagram, when learning new keypoints, we first locate both old and new keypoints on the diagram, as illustrated in Figure 2 (d), using their semantic definitions. We then iterate over each old keypoint to find the two closest newly defined keypoints, such as the orange keypoint in Figure 2 (d). Using these three keypoints, i.e., the orange one and the two green ones in Figure 2 (d), we construct the auxiliary prediction task for training the KA-Net.

This task creation can also be automated by GPT-4o as shown in Fig. 1. We only need to create the prompt to provide the name of the old keypoints that the model has learned previously and the name of the newly-defined keypoint. The prompt template is shown in Fig. 3. Empirically, we found that GPT-4o outputs the same task tuple as

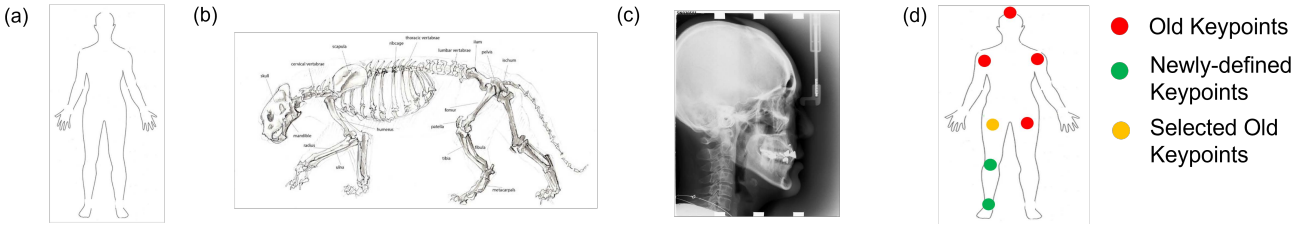

Figure 2. Illustration of the diagram for different objects, e.g., human and tiger body in (a) and (b), and the Cephalometric analysis [8] in (c). All the diagrams of each object category are readily found on the Internet. (d) is an example to demonstrate how we leverage the general diagram, e.g., the human skeleton diagram, to construct the auxiliary prediction task for training the KA-Net in IKL.

Content: Suppose I have defined the five keypoints

(**left hip, right hip, left shoulder, right shoulder, head**)

on the human body that the keypoint detector has learned.

Now, I have defined two new keypoints

(**left ankle, left knee**),

and I want to create an auxiliary learning task that uses these two new keypoints to predict the location of the most relevant keypoint and implicitly learn the association between the relevant old and new keypoint.

Which old keypoint should I choose to create the learning tuple and why? Please first answer fill in the blank in the format:

(left ankle, left knee, ?)

and then proceed to provide the reasoning

Figure 3. Illustration of the prompting template using the human body as an example to query GPT-4o to create this auxiliary task.

humans identified for all our experiments, making it a feasible solution to replace the human. The inference cost of prompting the GPT-4o is negligible.

5.4. Difference between Softmax Alternatives (main paper Sec. 3.2.2).

As mentioned in our main paper Sec. 3.2.2, We provide the visualization to demonstrate the difference between the Softmax alternatives, as shown in Fig. 4. Since most existing incremental learning (IL) literature uses image classification as the default visual task to evaluate the IL methods' performance, thus in methods like LWF [24] and its variants [5], when they calculate the negative log-likelihood between the old and new model's prediction for the old classes, they all perform the Softmax across difference old classes to obtain the normalized class prediction score. Such an operation in keypoint estimation is equivalent to normalizing each pixel location across each keypoint's heatmap prediction, as shown in Fig. 4 (a), which is a channel-wise normalization.

However, for the keypoint estimation task, it is always more critical to penalize the spatial-wise correctness for each keypoint individually [12, 19, 20, 22]. Thus in the present paper, given the task-oriented consideration, we explore the spatial-oriented knowledge distillation, where we perform the softmax over the heatmap spatial dimension, as shown in Fig. 4 (b), (c) and (d), where all these three operations are only normalized each pixel's value over each heatmap itself. Such a principle is also called instance-wise normalization. We will further provide the empirical study in the next section.

	5-Step Split MPII		
Method	AAA_4	AT_4	MT_4
LWF [24]	75.75	-3.86	0.41
KAMP (SM-2D)	78.54	0.52	1.78
KAMP (SM-Height)	79.26	0.96	3.01
KAMP (SM-Width)	79.35	1.33	3.22
KAMP (Ours, Eqn. 5)	**79.93**	**1.80**	**4.23**

Table 6. Ablation study of different Softmax alternatives

5.5. More Ablation Study of Different Softmax Alternatives (main paper Sec. 3.2.2)

In this section, we empirically explore whether those three spatial-wise Softmax alternatives differ in practice. The results are shown in Tab. 6, where SM-2D denotes the Softmax alternative as Fig.4 (d), SM-Height represents the Softmax alternative as Fig.4 (b), and SM-Width represents the Softmax alternative as Fig.4 (c), and Ours which averaging the SM-Width and SM-Height as defined in our main paper Eqn. (5).

We can observe that the SM-Width is slightly better than the other two alternatives. By further analysis, we empirically find that for the SM-2D, we need to calculate the exponential function for all the pixels over a certain heatmap

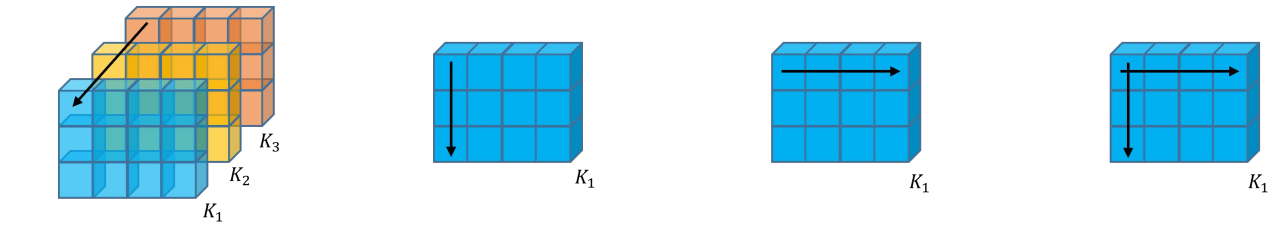

Figure 4. Illustration of different alternatives of Softmax operation.

	5-Step Split MPII		
Method	AAA_4	AT_4	MT_4
Ours (α=10)	78.45	0.21	0.58
Ours (α=1000)	79.14	0.33	2.97
Ours (α=100)	**79.93**	**1.80**	**4.23**

Table 7. Analysis of α of the Eqn (4) in the main paper

and summarize them as the denominator of the Softmax operator, which means we have total $H \times W$ terms in the denominator. Such a large denominator makes the value of each pixel small enough after the softmax, making the negative log-likelihood very small. This may further weaken the knowledge distillation effect since its scale will be too small and be less effective than the other two alternatives, i.e., SM-Height and -Width. For the SM-Height, we can see that in practice, it is only slightly worse than the SM-Width, and we conjecture that the spatial prior may be more readily consolidated via the width dimension. Finally, by combining the SM-Width and SM-Height as defined in our main paper Eqn. (5), we achieve the best overall performance.

6. Experimental Details and More Results.

6.1. More details of Evaluation Metrics (main paper Sec. 4)

As mentioned in our main paper, the Probability of Correct Keypoint (PCK) [27, 31, 37] is defined as PCK=$\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left(\frac{\|y_i-\hat{y}_i\|_2}{d}\leq\sigma\right)$, where a predicted keypoint location $\hat{y}_i$ is corrected if the normalized distance between $\hat{y}_i$ and the ground-truth location y_i over the longest side d of the ground-truth bounding box is less than the threshold σ. For the MPII and ATRW datasets, we use their defaulted σ as in [27, 29, 30].

For mean radial error (MRE) [38, 39], we adhere to the definition in previous studies [38, 39], MRE $= \frac{1}{N}\sum_{i}^{N}\sqrt{(x_i-\tilde{x}_i)^2+(y_i-\tilde{y}_i)^2}$, where $(\tilde{x}_i, \tilde{y}_i)$ denotes the predicted location of the keypoint while (x_i, y_i) denotes the ground-truth location. Since MRE measures the error between the prediction and ground-truth, a smaller value indicates better performance.

The definition of **A**verage **T**ransfer is: after step i, the average transfer over all previous steps is AT_i=$\frac{1}{i-1}\sum_{j=1}^{i-1}(a_{i,j}-a_{j,j})$, where $a_{i,j}$ denotes the average accuracy (PCK) or error (MRE) of the keypoints learned at step j after the training of step i. We time -1 to the AT_i when $a_{i,j}$ denotes error (MRE) since MRE is smaller the better. The definition of **M**aximal **T**ransfer is: MT_i=$\max_{k\in S_i}(R_{k,i}-\gamma_k)$, which represents the maximal performance change over all old keypoints after step i, where S_i denotes the set of old keypoints learned before step i, $R_{k,i}$ denotes the accuracy (PCK) or error (MRE) of the old keypoint k after step i, and γ_k denotes the initial accuracy (PCL) or error (MRE) of the old keypoint k when it was first learned in IKL. And we also time MT with -1 when MT measures the MRE.

6.2. More Dataset Statistics and Experimental Details. (main paper Sec. 4)

As mentioned in our main paper, we leverage the Head-2023 [4], Chest [16], MPII [2], and ATRW [21] datasets as our main testbed. For the Head-2023 dataset, we randomly chose 20 held-out images as our validation set, and randomly selected 80 held-out images as our test set. The Chest dataset only contains 279 training images after filtering by [41]. Given this small amount of training images, we randomly chose 79 held-out images as our test set without setting a validation set. Instead, we use the same training hyperparameter searched by the validation set of Head-2023 to directly use on the Chest dataset, except for the loss scale α as we use the same α scale as Split ATRW to ensure stable training. For the ATRW dataset, we randomly select 5% held-out images as our validation set and another 5% held-out image as our test set for the Split ATRW experiments, given the lack of public ground-truth annotations for the official ATRW test set.

For the Split MPII experiments, we leverage the SGD

optimizer, where we reduce the learning rate 10 times at 80 epochs. For Split Head-2023, Split Chest, and Split ATRW experiments, we follow the default optimizer for them, i.e., the Adam [17], where $\gamma_1 = 0.99$ and $\gamma_2 = 0$. As mentioned in our main paper, the total number of training epochs is 100 for all the methods. For our proposed method, we leverage 20 epochs for training the KA-Net in Stage-I, and we use the remaining 80 epochs for our Stage-II training, such that we can leverage the same number of epochs for both our method and other compared methods for a fair comparison.

6.3. More Details for the adaptation of low-shot regime. (main paper Sec. 4.3)

When our method needs to adapt for the low-data regime, we follow the same training strategy as [38] to train another auxiliary model in a self-supervised manner as [38] in the initial step (t=0). Specifically, for a given input image, we randomly select a point on the image and then randomly crop a patch containing that point. The same data augmentations as described in [38] are applied to this cropped image patch. The input image and the augmented image patch are then processed through two feature extractors, following the training approach of [38]. In the low-shot MPII experiment in Tab. 4 of our main paper, as we conventionally utilize the top-down human pose detector [30], we first identify each human instance in the image through detection. All detection results for the MPII dataset have been previously provided by top-down pose detectors like [9, 30]. Each detected human instance is then cropped from the image, and these cropped images are used as inputs for training the auxiliary model.

The training epoch for the auxiliary model is 100 and the same learning rate as in Sec. 4 of our main paper. We use the Adam optimizer with $\gamma_1 = 0.99$ and $\gamma_2 = 0$. Note that for the compared method CC2D [38] and EGT [39] they also have the same or similar self-supervised pertaining stage in their method to adapt for the low-shot regime for keypoint detection in medical images.

6.4. Standard Deviations of Tables 1 in Our Main Paper (main paper Sec. 4).

As described in our main paper's Sec. 4, we report the per-step results and standard deviations of each dataset in Table 1 of our main paper in Tab. 11, Tab. 12, Tab. 13 and 14, respectively, where we observe that our method enjoys relatively more minor deviation and consistently outperform all the comparison methods for all the datasets.

6.5. Analysis of α (main paper Sec. 4)

As mentioned in our main paper's Sec. 4, in Tab. 7, we further provide more analysis of the hyperparameter α in Eqn. (4) in our main paper. For all our experiments in the main paper and supplementary, we use the held-out validation set to determine the α, the same strategy as how we determine the hyperparameters of all the comparison methods. Generally, when the α is large, the knowledge consolidation may dominate the Stage-II training in our proposed method, and thus the acquisition of the new keypoint may be hindered. While if the α is too small, the old keypoint may be forgotten catastrophically in IKL. Therefore, there should exist a proper α that can achieve a good balance between the new keypoint acquisition and avoid the catastrophic forgetting of the old keypoints. Results in Tab. 7 empirically support our analysis, where we can see that when α increases 10 times from 10 to 1000, the proper α is over 100 (1e2) that can achieve the proper average performance (AAA_4).

6.6. Details of the Keypoint Group. (main paper Sec. 4)

For the Head-2023 dataset, there are 38 keypoints: Nasal root, Nasal bridge, Outer canthus, Inner canthus, Upper incisal tip, Lower incisal tip, Chin tip, Anterior chin, Inferior chin, Posterior chin, Lower anterior tooth plane, Upper anterior incisal point, Superior protuberance, Inferior protuberance, Lower jaw point, Pre-molar anterior chin, Posterior nostril, Anterior nostril, Beauty midpoint, Center point of the mandibular posterior platform, Upper central incisor tip, Starting point of the mandibular incisor, Small lip spicy point, Red point above the nose, Chin apex, Chin apex, Flat base point, PT point, Bolton point, Upper lip fine point, Lower lip fine point, Alveolar anterior chin point, Alveolar inferior chin point, Chin point, Alveolar chin root point, Chin point, Upper lip external point, Lower lip external point. As mentioned in our main paper Sec. 4, we select the first 19 keypoints as our first group and then splits the rest of the keypoints into 4 groups, where we randomly select two or more keypoints for each incremental step. For the Chest dataset [16, 41], it contains six keypoint as the top, the bottom, and the right boundary point of the right lung and the same three keypoints for the left lung.

For the MPII dataset, there are 16 human body keypoints: right ankle, right knee, right hip, left hip, left knee, left ankle, pelvis, thorax, upper neck, head top, right wrist, right elbow, right shoulder, left shoulder, left elbow and left wrist. We randomly select five keypoints for the initial step, i.e., Step-0, and then we randomly select two or more keypoints for each incremental step. The ATRW dataset has 15 keypoints of Amur Tiger: left ear, right ear, nose, right shoulder, right front paw, left shoulder, left front paw, right hip, right knee, right back paw, left hip, left knee, left back paw, tail, and center. We randomly choose 6 keypoints for the Step-0 of Split ATRW and then randomly choose two or more keypoints for each incremental. For instance, for the qualitative results shown in our main paper and the supplementary, we chose the upper neck, left elbow, right wrist, right knee, and left ankle for the Step-0 training of

	Step-1		
Method	AAA_1	AT_1	MT_1
EWC [18]	67.13	-19.69	0.14
RW [6]	59.12	-14.97	-9.62
MAS [1]	72.12	-5.27	0.44
LWF [24]	77.76	0.31	1.81
AFEC [32]	68.18	-4.16	-0.55
CPR [5]	77.08	0.49	1.33
KAMP (Ours)	**79.17**	**1.94**	**3.34**

Table 8. Result of the balanced setup

Split MPII; and we selected nose, tail, right back paw, left back paw, right front paw, and left front paw as the keypoint group introduced in Step-0 for the Split ATRW.

6.7. More Results of Another Setup: Balanced Number of Old and New Keypoints.

For all the experiments before this section, in each incremental step, the number of old keypoints introduced previously is always larger than the number of new keypoints introduced in the current step. Under such a setting, methods that can well preserve the performance of old keypoints will outperform others since the performance of old keypoints may dominate the average accuracy metric, i.e., AAA.

To provide a more comprehensive view of our method, in this section, we consider a balanced setup where only one incremental step is introduced, and the number of new keypoints is the same as the old ones. In such a balanced setup, we can further see whether our method still has superiority over other methods. As shown in Tab. 8, we can see that the gap between each comparison method is smaller than the gap we observed in our previous experiments. Our proposed method still achieve the largest average accuracy (AAA) positive average transfer (AT) among all the other methods. This further demonstrates that the superiority of our method is general and consistent over different experimental setups.

6.8. More Results of Another Setup: Old Keypoints Only for the Upper Body, New Keypoints only for the Lower Body in Split MPII.

To further explore whether our proposed KAMP method can consistently perform well under different setups, here we use the MPII dataset to create a 2-Step protocol where the old keypoints are all from the upper body of the human while the newly-defined keypoints are all from the lower body. Such a scenario can be viewed as a kind of "extrapolation" as the keypoints of the lower body are all outside of the upper body. There are not so many physical connections between the upper body and the lower body, and thus the locations of the keypoints of the lower body may not be highly related to the keypoints in the upper body. Therefore such a protocol would be much more challenging than our previous protocols. As shown in Tab. 9, compared with the competitive method, CPR [5], our proposed method can still achieve positive average transfer and maximal transfer under this challenging protocol and also outperform CPR with a large margin on the average performance. The absolute value of the average transfer and maximal transfer for our method is small, which is expected as explained above. However, our method as a novel baseline for IKL still demonstrates its superiority, and it is also promising for us to explore better methods to further boost the performance in the future.

	AAA_1	AT_1	MT_1
CPR [5]	81.48	-5.33	-0.42
Ours	**84.53**	**0.04**	**0.63**

Table 9. Experimental results when we first learn the keypoints all from the upper body of the human and then incrementally learn the new keypoints all from the lower body using the MPII dataset.

6.9. More Experimental Details about the Low-shot Experiments between our IKL and other Alternative Methods. (main paper line 580-581)

Here we provide more experimental details about the low-shot experiments between our proposed IKL setting and other alternative methods of our Sec. 4.3 in our main paper. For the experiment on Head-2023, since both CC2D [38] and EGT [39] have not trained on Head-2023, thus we use the official implementation of them to run the experiment on Head-2023 to get the results. All the training details are the same as their original paper [38, 39].

For the unsupervised keypoint learning (UKL) [13] in Split MPII, we choose the SOTA UKL method [13] to perform our experiments. We follow the same experimental details in [13] to conduct the unsupervised pertaining on the MPII datasets, where we pre-define the model to output 32 keypoints without assigning any semantic meaning for each one. Then after the unsupervised pretraining on each dataset, we follow the standard practice in UKL [13, 26, 28, 40] that we freeze the unsupervised pretrained model and then learn a linear transformation between the pre-defined keypoints and each newly-defined keypoint introduced in each incremental step.

For the category-agnostic pose estimation (CAPE) [7] in Split MPII, we leverage its SOTA method, i.e., MetaPoint+ [7], to conduct our experiments. We also follow the same experimental details in [7], where we treat the keypoint categories related to humans (e.g., human body, face, and hand) as the unseen category to avoid information leakage when performing the pertaining in CAPE. This is similar to the cross super-category experiments in Sec. 4.3 in

[34]. The difference is that in our experiments, we need to perform the testing on each keypoint group incrementally.

6.10. Per-keypoint Performance of Our Method under the ATRW. (main paper Sec. 4.1)

As mentioned in our main paper's Sec. 4.1, here we report the per-keypoint's performance of knowledge transfer after three incremental steps of Split ATRW. As shown in Tab. 10, 6 over 13 old keypoints have non-negative transfer after three incremental steps. This further verifies our conjecture that there is much positive transfer occurring for many old keypoints to **offset** the forgetting of other old keypoints such that our method can achieve a very small negative average transfer.

6.11. More Visualization Results. (main paper Fig. 4)

As mentioned in our main paper's Fig. 4, we include more visualization results in Fig. 5. Again, our method can achieve more structurally correct keypoint prediction and less miss-detection error than other comparison methods.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018. 9, 11, 12

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7

[3] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 455–472. Springer, 2020. 3

[4] Jun Cao, Juan Dai, Xuguang Li, Bingsheng Huang, Ching-Wei Wang, and Hongyuan Zhang. Cephalometric landmark detection in lateral x-ray images, 2023. 7

[5] Sungmin Cha, Hsiang Hsu, Taebaek Hwang, Flavio Calmon, and Taesup Moon. {CPR}: Classifier-projection regularization for continual learning. In *International Conference on Learning Representations*, 2021. 6, 9, 11, 12

[6] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–547, 2018. 9, 11, 12

[7] Junjie Chen, Jiebin Yan, Yuming Fang, and Li Niu. Metapoint learning and refining for category-agnostic pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23534–23543, 2024. 9

[8] Runnan Chen, Yuexin Ma, Nenglun Chen, Daniel Lee, and Wenping Wang. Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 873–881. Springer, 2019. 6

[9] 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. 8

[10] Jiahua Dong, Wenqi Liang, Yang Cong, and Gan Sun. Heterogeneous forgetting compensation for class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11742–11751, 2023. 11, 12

[11] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4040–4050, 2021. 11, 12

[12] Kerui Gu, Linlin Yang, and Angela Yao. Removing the bias of integral pose regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11067–11076, 2021. 6

[13] Xingzhe He, Bastian Wandt, and Helge Rhodin. Autolink: Self-supervised learning of human skeletons and object outlines by linking keypoints. In *Advances in Neural Information Processing Systems*, 2022. 2, 9

[14] Zhongzhan Huang, Senwei Liang, Mingfu Liang, and Haizhao Yang. Dianet: Dense-and-implicit attention network. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4206–4214, 2020. 3

[15] Zhongzhan Huang, Senwei Liang, and Mingfu Liang. A generic shared attention mechanism for various backbone neural networks. *Neurocomputing*, 611:128697, 2025. 3

[16] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, George Thoma, Yi-Xiang Wang, Pu-Xuan Lu, and Clement J McDonald. Automatic tuberculosis screening using chest radiographs. *IEEE Trans. Med. Imaging*, 33(2):233–245, 2014. 7, 8

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8

[18] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 9, 11, 12

[19] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11025–11034, 2021. 6

	KP1	KP2	KP3	KP4	KP5	KP6	KP7	KP8	KP9	KP10	KP11	KP12	KP13
Transfer	1.13	-2.25	-2.86	-14.62	-1.71	1.13	2.92	-0.58	5.13	-3.09	2.47	-1.41	0.00

Table 10. Per-keypoint performance transfer in Split ATRW after 3 incremental steps. KP is the abbreviation of keypoint.

Method	Step-1			Step-2			Step-3			Step-4		
	A-MRE$_1$ ↓	AT$_1$	MT$_1$	A-MRE$_2$ ↓	AT$_2$	MT$_2$	A-MRE$_3$ ↓	AT$_3$	MT$_3$	A-MRE$_4$ ↓	AT$_4$	MT$_4$
EWC [18]	4.12±0.97	-1.2±0.19	0.02±0.14	6.36±1.83	-2.98±0.65	-0.87±0.38	8.86±2.73	-5.76±1.87	-2.78±0.99	10.97±2.76	-6.37±2.14	-4.76±1.53
RW [6]	3.67±0.38	-0.57±0.32	0.28±0.07	4.25±0.69	-1.65±0.39	-0.43±0.37	5.75±0.65	-2.53±0.76	-0.91±0.47	6.49±1.73	-4.23±1.12	-2.88±0.85
MAS [1]	3.87±0.10	-0.76±0.27	0.14±0.15	4.37±0.53	-1.12±0.65	-0.05±0.18	4.87±0.93	-1.59±0.62	-0.28±0.81	5.31±0.54	-2.15±0.37	-1.33±0.38
LWF [24]	3.06±0.04	0.06±0.07	0.27±0.06	3.45±0.39	-0.43±0.24	0.21±0.29	3.99±0.25	-1.34±0.41	-0.39±0.52	4.31±0.26	-1.26±0.20	0.57±0.25
AFEC [32]	3.12±0.05	-0.04±0.05	0.42±0.12	3.96±0.65	-0.84±0.16	0.05±0.38	4.94±1.02	-1.65±0.87	-0.81±0.64	5.77±0.94	-3.45±0.69	-1.46±0.75
CPR [5]	2.96±0.02	0.29±0.03	0.72±0.06	3.18±0.12	-0.06±0.28	0.33±0.11	3.46±0.16	-0.76±0.12	0.63±0.08	3.71±0.41	-1.18±0.30	0.16±0.13
SFD [11]	3.45±0.03	0.12±0.11	0.02±0.06	3.59±0.11	0.01±0.05	0.13±0.02	4.43±0.07	-0.18±0.06	0.19±0.02	4.76±0.11	-0.43±0.07	0.02±0.04
WF [33]	3.37±0.02	0.14±0.03	0.35±0.11	3.50±0.13	0.03±0.02	0.19±0.03	4.29±0.10	-0.01±0.11	0.21±0.16	4.58±0.23	0.03±0.09	0.11±0.03
GBD [10]	3.28±0.04	0.21±0.10	0.43±0.07	3.41±0.09	0.02±0.05	0.10±0.03	4.18±0.08	0.02±0.03	0.27±0.08	4.34±0.17	0.12±0.08	0.47±0.02
KAMP (Ours)	**2.13±0.04**	**0.63±0.08**	**0.92±0.12**	**2.25±0.07**	**0.36±0.06**	**0.72±0.19**	**2.29±0.03**	**0.33±0.02**	**0.65±0.07**	**2.32±0.09**	**0.41±0.03**	**0.84±0.09**

Table 11. Results on Split Head-2023 after 5 Step IKL, starting from the same Step-0 trained model. A-MRE: smaller the better

Method	Step-1		
	A-MRE$_1$ ↓	AT$_1$	MT$_1$
EWC [18]	13.28±2.31	-8.23±1.87	-3.67±1.21
RW [6]	9.48±1.53	-7.12±1.29	-4.15±0.76
MAS [1]	7.36±1.02	-1.86±0.83	-0.19±0.64
LWF [24]	6.35±0.09	-1.34±0.12	0.18±0.07
AFEC [32]	8.04±0.28	-2.67±0.31	0.15±0.0.10
CPR [5]	6.17±0.06	-0.87±0.03	0.29±0.05
SFD [11]	7.68±0.03	-0.54±0.04	0.13±0.01
WF [33]	7.31±0.03	-0.31±0.02	0.16±0.04
GBD [10]	6.42±0.02	0.06±0.07	0.21±0.03
KAMP (Ours)	**5.67±0.08**	**0.29±0.03**	**0.62±0.09**

Table 12. Results on Split Chest after 2 Step IKL, starting from the same Step-0 trained model. A-MRE: smaller the better

[20] Jiefeng Li, Tong Chen, Ruiqi Shi, Yujing Lou, Yong-Lu Li, and Cewu Lu. Localization with sampling-argmax. *Advances in Neural Information Processing Systems*, 34: 27236–27248, 2021. 6

[21] Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. Atrw: A benchmark for amur tiger re-identification in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2590–2598, 2020. 7

[22] Yanjie Li, Sen Yang, Shoukui Zhang, Zhicheng Wang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Is 2d heatmap representation even necessary for human pose estimation? *arXiv preprint arXiv:2107.03332*, 2021. 6

[23] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11313–11322, 2021. 3

[24] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 3, 4, 5, 6, 9, 11, 12

[25] Senwei Liang, Zhongzhan Huang, Mingfu Liang, and Haizhao Yang. Instance enhancement batch normalization: An adaptive regulator of batch noise. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4819–4827, 2020. 3

[26] Dimitrios Mallis, Enrique Sanchez, Matthew Bell, and Georgios Tzimiropoulos. Unsupervised learning of object landmarks via self-training correspondence. *Advances in Neural Information Processing Systems*, 33:4709–4720, 2020. 9

[27] Olga Moskvyak, Frederic Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Semi-supervised keypoint localization. In *International Conference on Learning Representations*, 2021. 7

[28] Enrique Sanchez and Georgios Tzimiropoulos. Object landmark discovery through unsupervised adaptation. *Advances in Neural Information Processing Systems*, 32, 2019. 9

[29] 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 (CVPR)*, pages 7308–7317, 2023. 2, 7

[30] 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. 7, 8

[31] Can Wang, Sheng Jin, Yingda Guan, Wentao Liu, Chen Qian, Ping Luo, and Wanli Ouyang. Pseudo-labeled auto-curriculum learning for semi-supervised keypoint localization. In *International Conference on Learning Representations*, 2022. 7

[32] Liyuan Wang, Mingtian Zhang, Zhongfan Jia, Qian Li, Chenglong Bao, Kaisheng Ma, Jun Zhu, and Yi Zhong. Afec: Active forgetting of negative transfer in continual learning. *Advances in Neural Information Processing Systems*, 34: 22379–22391, 2021. 9, 11, 12

[33] Jia-Wen Xiao, Chang-Bin Zhang, Jiekang Feng, Xialei Liu, Joost van de Weijer, and Ming-Ming Cheng. Endpoints weight fusion for class incremental semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer*

	Step-1			Step-2			Step-3			Step-4		
Method	AAA_1	AT_1	MT_1	AAA_2	AT_2	MT_2	AAA_3	AT_3	MT_3	AAA_4	AT_4	MT_4
EWC [18]	52.42±4.68	-26.18±6.6	-0.95± 0.01	32.37± 5.8	-50.08± 6.61	-25.53± 11.27	24.67±1.75	-57.59± 2.82	-35.56± 12.02	38.64± 0.40	-51.84±0.95	-12.21± 2.69
RW [6]	58.85±0.79	-3.19±2.91	-0.41±0.007	51.12±1.78	-6.88±0.79	-0.66±0.14	43.44±0.05	-8.20±6.36	-0.89±4.46	38.47±2.65	-18.83±0.51	-7.13±3.72
MAS [1]	65.35±6.14	-6.97±7.06	-0.24±0.10	59.99±2.00	-13.85±2.59	-0.10±0.07	63.29±2.89	-9.91±3.7	0.00±0.17	67.03±1.65	-7.56±0.7	0.34±0.20
LWF [24]	71.23±0.46	-0.49±0.35	1.41±0.16	73.75±0.65	-1.03±0.60	0.28±0.17	74.69±0.56	-1.63±0.38	0.68±0.11	75.75±0.51	-3.86±1.76	0.41±0.31
AFEC [32]	63.54±0.79	-4.78±0.44	-1.50±0.41	52.98±3.2	-17.26±2.37	-7.33±1.87	47.25±8.18	-18.30±11.32	-6.17±8.77	37.24±0.05	-22.85±1.06	-15.42±0.07
CPR [5]	71.67±0.11	0.68±0.04	2.17±0.08	72.86±0.42	-2.42±0.30	0.61±0.16	73.95±0.36	-2.07±0.55	1.19±0.30	75.52±0.67	-3.24±1.00	0.75±0.62
SFD [11]	68.79±0.16	0.07±0.01	0.53±0.12	70.09±0.35	-0.98±0.33	0.71±0.12	71.15±0.38	-0.41±0.28	0.85±0.05	71.49±0.82	-0.93±0.52	0.21±0.08
WF [33]	69.92±0.19	0.08±0.13	0.76±0.06	70.94±0.29	-1.31±0.26	0.58±0.18	72.24±0.36	-1.65±0.07	1.54±0.14	72.87±0.39	-0.46±0.04	0.38±0.14
GBD [10]	71.94±0.06	0.18±0.06	1.76±0.09	72.69±0.26	-1.11±0.17	0.69±0.05	74.74±0.67	-0.10±0.38	0.64±0.18	75.62±0.18	-0.18±0.06	0.35±0.19
KAMP (Ours)	**73.54±0.34**	**2.23±0.37**	**3.98±0.25**	**76.55±0.11**	**2.30±0.04**	**3.68±0.02**	**78.06±0.49**	**2.24±0.35**	**4.12±0.33**	**79.93±0.12**	**1.80±0.11**	**4.23±0.31**

Table 13. Results on Split MPII after 5 Step IKL, starting from the same Step-0 trained model.

	Step-1			Step-2			Step-3		
Method	AAA_1	AT_1	MT_1	AAA_2	AT_2	MT_2	AAA_3	AT_3	MT_3
EWC [18]	40.33±4.06	-92.58±8.81	-81.71±3.76	25.19±2.63	-62.88±6.57	-17.28±6.75	14.38±1.01	-59.75±8.13	-2.08±1.6
RW [6]	82.90±0.32	-1.09±1.14	-0.56±0.61	84.14±1.63	-8.10±2.95	0.00±1.26	84.15±0.43	-10.87±2.61	0.00±1.12
MAS [1]	89.54±0.4	-7.26±0.45	-0.56±1.12	88.26±0.06	-5.80±0.22	-0.62±1.18	85.68±1.43	-5.80±4.31	-1.13±0.57
LWF [24]	90.45±0.19	-6.18±1.06	-4.52±1.69	89.15±0.61	-4.99±0.63	2.47±0.74	87.31±1.05	-5.10±2.83	-0.64±1.21
AFEC [32]	61.57±1.52	-30.46±0.05	-10.11±0.69	45.70±0.52	-35.75±1.16	-9.26±0.98	33.03±0.75	-40.25±0.88	-8.02±0.15
CPR [5]	90.86±0.36	-3.24±0.09	0.00±0.69	90.43±0.46	-2.22±1.39	1.85±0.78	89.34±0.73	-2.76±0.75	4.49±0.12
SFD [11]	88.98±0.36	-2.19±0.11	-0.42±0.69	87.54±0.47	-1.88±1.09	1.18±0.75	86.11±0.81	-1.13±0.21	0.41±0.98
WF [33]	90.16±0.24	-2.08±0.02	-0.35±0.52	88.63±0.38	-1.76±0.96	1.77±0.68	86.69±0.87	-0.97±0.29	0.62±0.54
GBD [10]	90.86±0.36	-3.24±0.09	0.00±0.69	89.03±0.18	-1.23±0.67	1.84±0.98	87.42±0.77	-0.89±0.30	0.65±0.25
KAMP (Ours)	**93.21±0.76**	**-0.86±0.27**	**0.56±0.32**	**93.63±0.24**	**-0.34±0.21**	**3.08±0.61**	**93.16±0.34**	**-0.84±0.28**	**5.13±0.47**

Table 14. Results on Split ATRW after 4 Step IKL, starting from the same Step-0 trained model.

Vision and Pattern Recognition, pages 7204–7213, 2023. 11, 12

[34] 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. 10

[35] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 3

[36] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11802–11812, 2021. 3

[37] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11802–11812, 2021. 7

[38] Qingsong Yao, Quan Quan, Li Xiao, and S Kevin Zhou. One-shot medical landmark detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 177–188. Springer, 2021. 2, 4, 7, 8, 9

[39] Zihao Yin, Ping Gong, Chunyu Wang, Yizhou Yu, and Yizhou Wang. One-shot medical landmark localization by edge-guided transform and noisy landmark refinement. In *European Conference on Computer Vision*, pages 473–489. Springer, 2022. 2, 7, 8, 9

[40] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018. 9

[41] Heqin Zhu, Qingsong Yao, Li Xiao, and S Kevin Zhou. You only learn once: Universal anatomical landmark detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 85–95. Springer, 2021. 7, 8

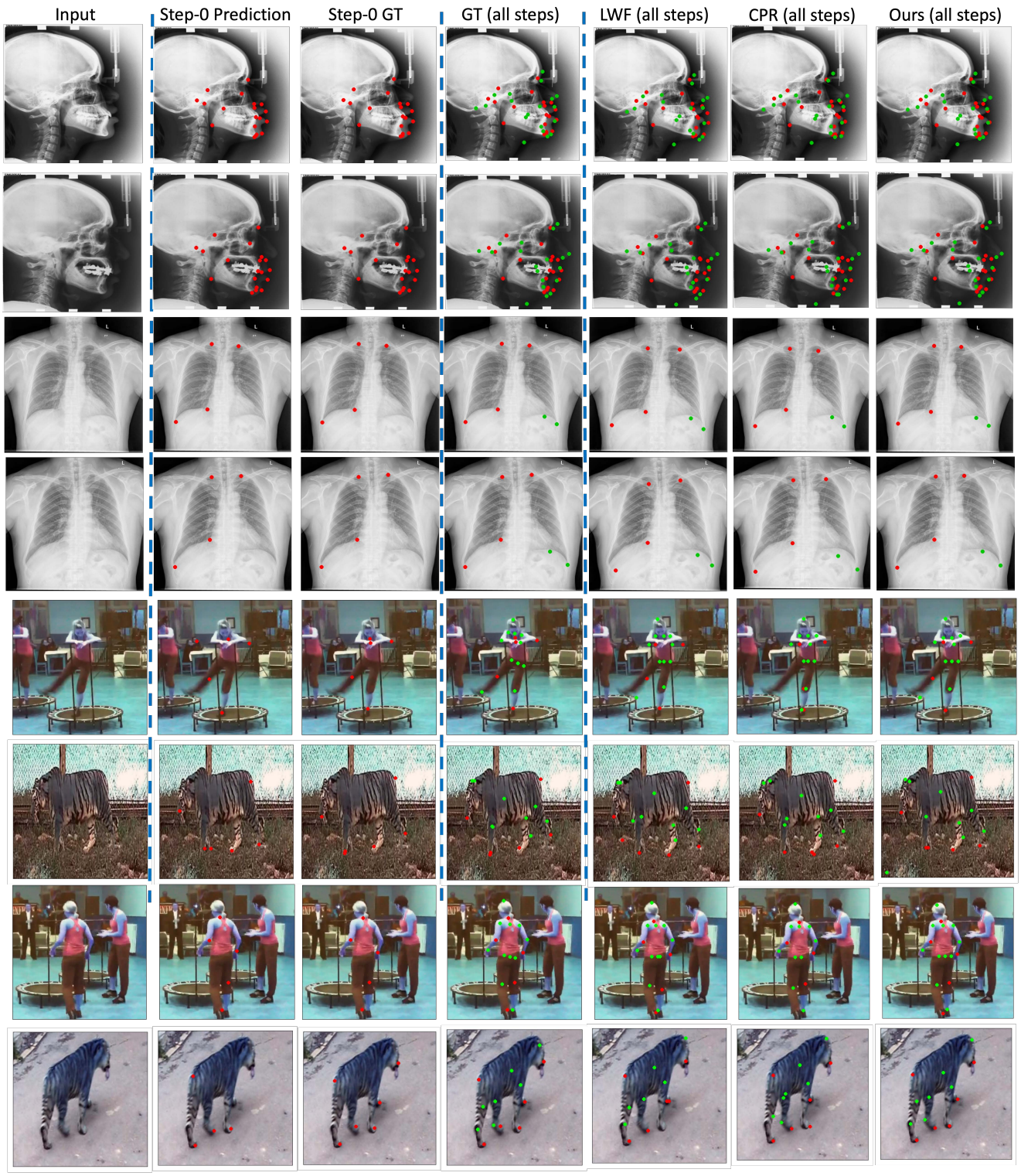

Figure 5. Visualization results on Split Head-2023, Split Chest, Split MPII and Split ATRW. All the methods start from the same Step-0 model, whose prediction is shown in the second column. GT denotes ground truth. The red circles denote the keypoints learned in Step 0, while the green circles denote all the new keypoints learned in later steps. We observe that after the IKL, the compared methods (LWF and CPR) may acquire the new keypoints as ours, but they have obvious miss-detection and wrong estimation (e.g., out of the body). While our method can consistently associate the new and old keypoints and achieve structurally accurate keypoint predictions.