Incremental Object Keypoint Learning

Mingfu Liang[1] Jiahuan Zhou[2,*] Xu Zou[3] Ying Wu[1]
[1] Northwestern University [2] Wangxuan Institute of Computer Technology, Peking University
[3] Huazhong University of Science and Technology

Abstract

*Existing progress in object keypoint estimation primarily benefits from the conventional supervised learning paradigm based on numerous data labeled with pre-defined keypoints. However, these well-trained models can hardly detect the undefined new keypoints in test time, which largely hinders their feasibility for diverse downstream tasks. To handle this, various solutions are explored but still suffer from either limited generalizability or transferability. Therefore, in this paper, we explore a novel keypoint learning paradigm in that we only annotate new keypoints in the new data and incrementally train the model, without retaining any old data, called **I**ncremental object **K**eypoint **L**earning (IKL). A two-stage learning scheme as a novel baseline tailored to IKL is developed. In the first Knowledge Association stage, given the data labeled with only new keypoints, an auxiliary KA-Net is trained to automatically associate the old keypoints to these new ones based on their spatial and intrinsic anatomical relations. In the second Mutual Promotion stage, based on a keypoint-oriented spatial distillation loss, we jointly leverage the auxiliary KA-Net and the old model for knowledge consolidation to mutually promote the estimation of all old and new keypoints. Owing to the investigation of the correlations between new and old keypoints, our proposed method can not just effectively mitigate the catastrophic forgetting of old keypoints, but may even further improve the estimation of the old ones and achieve a positive transfer beyond anti-forgetting. Such an observation has been solidly verified by extensive experiments on different keypoint datasets, where our method exhibits superiority in alleviating the forgetting issue and boosting performance while enjoying labeling efficiency even under the low-shot data regime.*

1. Introduction

As a fundamental task in computer vision, estimating the visual keypoint locations of an object serves as the indispensable primitive to support numerous down-stream tasks, e.g., object pose detection [29, 77, 80], tracking [28], action recognition [7], generation [57] and animation [28, 58], etc. Over a long period of time, **supervised keypoint learning (SKL)** [19, 36, 39, 68] has significantly advanced the progress on keypoint estimation. By exploiting large-scale pure 2D keypoint datasets [5, 44] with extreme variance, SKL aims to train a deep neural network that can robustly detect a set of pre-defined keypoints in test images.

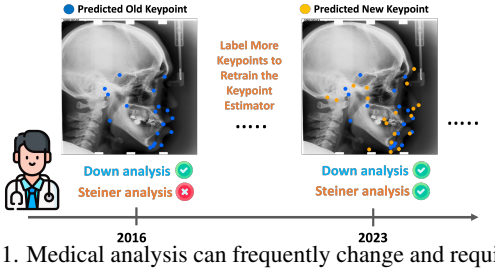

Figure 1. Medical analysis can frequently change and require new keypoints essentially [6], while the labeling is time-consuming.

However, the SKL models can not estimate newly-added undefined keypoints of an object, while new demands from downstream tasks will inevitably require new keypoints. For instance, until now, models trained by existing keypoint dataset [67] can only detect 19 keypoints to support Downs analysis in the Cephalometric analysis, while they are infeasible for important analysis like Steiner analysis [53], as shown in Fig. 1. Such an issue motivated a recent MICCAI challenge [6] on annotating much more keypoints to support increasing analysis. However, keypoint labeling is time-consuming [18, 45, 50], especially when keypoints are defined sequentially. Another alternative solution is to train separate new keypoint estimator [12, 31] for each set of new keypoints as in Fig. 2. However, the number of estimators will increase linearly, and the separate training strategy will make each estimator easily overfit the new keypoints [12] and hardly capture the intrinsic relation between old and new keypoints, leading to sub-optimal performance.

To mitigate the above issues, **unsupervised keypoint learning (UKL)** [22, 23, 49, 55, 84] is proposed for new keypoint estimation. The models are firstly pretrained in an unsupervised manner to predict numbers of keypoints ran-

*: Corresponding author

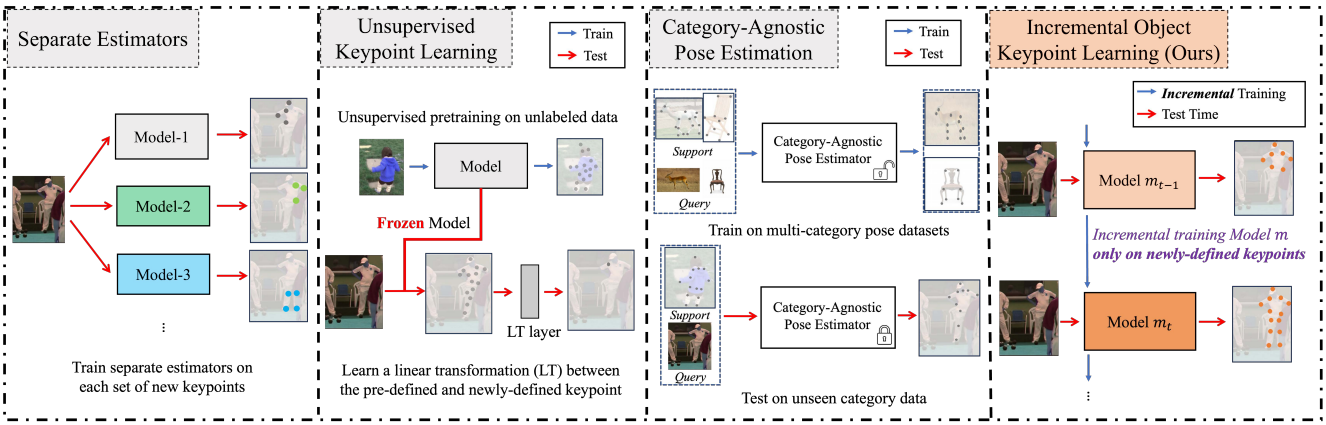

Figure 2. Separate Estimators need multiple estimators in test time. Unsupervised keypoint learning (UKL) and category-agnostic pose estimation (CAPE) exploit a fixed pretrained keypoint estimator on unseen new keypoints. While our Incremental object Keypoint Kearning (IKL) continually updates the same model on the new data labeled **only on new keypoints without retaining any old training data**.

domly, then a linear transformation is learned to transfer the existing keypoints to the new undefined ones. However, UKL is merely effective on objects with relatively rigid motion or video data with consistent object instances and local changes in appearance and motion. As a result, it is hard to perform UKL on large-scale pure 2D datasets [5, 44] due to their extreme inter-image differences in the background, appearance, and motions of distinct object instances. Recently, a supervised metric learning-based [59] paradigm named **category-agnostic pose estimation (CAPE)** [76] is proposed to learn a class-agnostic pose estimator that can estimate unseen keypoints in test time with a few labeled support samples. However, as the model is frozen after pretraining, CAPE is limited by its restricted generalization and transferability for newly-defined keypoints (Tab. 4).

Therefore, in this paper, we propose to solve the new keypoint estimation problem creatively in an incremental learning (IL) manner. As shown in Fig. 2, we propose an **Incremental object Keypoint Learning (IKL)** paradigm to incrementally train a keypoint estimator on the **new data labeled only with new keypoints**. Even without retaining any old data, the IKL model should not forget the old keypoints catastrophically. Compared to the naive solutions that either label both new and old keypoints or learn separate estimators, our IKL only labels the new keypoints and maintains *just one model in the lifetime*, which is computationally efficient. Moreover, different from the UKL and CAPE whose performances are largely limited by their restricted transferability depending on a fixed pretrained model, the IKL alleviates such an issue by continually training the model to sufficiently expand its knowledge on newly-defined keypoints and may *even improve the old keypoint estimation* when learning more relevant new keypoints. Lastly, as most works in IL mainly study the classification task [13, 70], to our best knowledge, IKL as an incremental keypoint regression task has rarely been studied before, which also provides novel insights for IL.

However, IKL also poses a new challenge known as label non-co-occurrence (LNCO). As only new keypoints are labeled in the new data, the kinematic and anatomical constraints between old and new keypoints are not explicitly presented in the label space, making it hard for the model to capture such a physical prior during incremental training. Existing IL methods (Tab. 1) do not explicitly model such inter-keypoint relations in IKL, thus the old and new keypoint predictions can not mutually support each other and may readily bias toward the new ones, exacerbating the forgetting of the old keypoints (Tab. 1).

To tackle the challenge of IKL, we propose a novel two-stage learning scheme called **KAMP** as a new baseline tailored to IKL. In the first **K**nowledge **A**ssociation stage, we train an auxiliary KA-Net to associate old and new keypoints based on their physiological connection, represented by their spatial adjacency. Specifically, KA-Net learns to predict the selected old keypoint given the related new keypoints to acquire their intrinsic anatomical relevance. In the second **M**utual **P**romotion stage, we train the new model on all keypoints to mutually promote their estimation, where the old ones are updated by distilling from the KA-Net and the old model with a newly designed keypoint-oriented spatial distillation loss for better knowledge preservation. To verify our effectiveness, we simulate IKL based on four keypoints datasets on both medical and natural images, i.e., Cephalometric [6], Chest [27], MPII [5] and ATRW [37]. Extensive results demonstrate that our KAMP can not just effectively alleviate the catastrophic forgetting of old keypoints, but even further boost their performance and thus outperform the existing exemplar-free IL methods by a significant margin (Tab. 1) and can also work for low-shot scenarios (Tab. 3 and 4). Our analysis further reveals that IKL is complementary to learning paradigms like CAPE (Tab. 4) and labeling-efficient for new keypoint estimation.

To summarize, our contributions are three-fold: (1) We establish a novel paradigm, Incremental object Keypoint Learning (IKL), to tackle the challenging demands of new keypoints estimation. (2) We propose a two-stage exemplar-free IKL method to explicitly model and exploit the relation between old and new keypoints to help alleviate the label non-co-occurrence problem tailored to IKL, which may further boost the performance of old keypoints beyond anti-forgetting. (3) As a proof-of-concept, we empirically show that IKL is practical and labeling-efficient to scale up a pretrained keypoint estimator on new keypoint estimation, which is much better than other alternative paradigms.

2. Related Works

Keypoint Estimation (KE). Existing research of KE focuses on estimating a fixed number of keypoints for a specific category [5, 29, 44, 54, 78–80]. Most works design new methods [20, 34, 35, 38, 89] or network architectures [39, 48, 61, 62, 68, 77] to improve the supervised learning over large-scale pure 2D keypoint datasets [5, 44]. Estimating dense keypoints [29, 82] of the human body has been proposed but requires extreme labeling cost. Recent works explore semi-supervised [24, 50, 66] or unsupervised learning [63, 84] to reduce the laborious labeling consumption, but the unsupervised methods still hardly achieve a precise estimation of semantically meaningful keypoints [49]. Recently, Category-agnostic keypoint estimation (CAPE) [10, 56, 76] is proposed to learn a class-agnostic model to detect keypoints of unseen categories by a few labeled support images without retraining the model. However, as the model is frozen after pretraining, CAPE is limited by restricted generalization and transferability due to the intra-class appearance variation, self-occlusion, and appearance ambiguity, as detailed in [76]. Different from the above paradigm that pre-defined a set of keypoints or assumed the pretrained model was sufficient enough for different downstream tasks, our proposed IKL paradigm continuously updates the model only on the new keypoints to increasingly expand its knowledge, which is much more flexible and labeling-efficient.

Incremental Learning (IL). Existing IL methods can be categorized as exemplar-free and exemplar-based ones [3, 25, 26, 32, 41, 52], given whether old data can be retained. However, saving exemplars not just introduces increasing costs on memory, but also has privacy issues in real-world applications. For exemplar-free methods, existing approaches focus on adding regularization over parameters [2, 33, 47, 81, 83] or the network's output in different positions [15, 40, 73, 85]. Regarding the IL settings, adding new classes [13] or domains [41, 46, 65] have been studied popularly and mainly tested on classification benchmarks [13, 14]. In contrast, keypoint estimation is a location regression task [20, 34, 48], and our proposed IKL aims to detect newly defined keypoints incrementally. The incremental animal pose estimation [51] (IAPE) is a related setting to us. However, IAPE assumes the label space of keypoints is **fixed** for each animal, and it only learns new animals' poses, i.e., only the input distribution is changed with new animal domains. While for our IKL, we instead consider **adding new keypoints** to the label space for an object category, and the old and new data distribution may be relatively correlated.

Positive Transfer (PT) in IL. Most IL approaches focus on alleviating the catastrophic forgetting problem [13], while only a few works are dedicated to achieving positive transfer (PT) in IL [30, 43, 47], i.e., improving the old tasks while learning a new task. Those methods encourage PT by modifying the gradient updates based on the new and old tasks' correlations which are measured by saving the data [47] from the old task or based on formal analyses [30, 43]. Differently, our KAMP achieves PT by explicitly **modeling the relation between old and new keypoints** based on their intrinsic anatomical relevance, and we **exploits the relation to design novel distillation to further boost the estimation of the old keypoint**, which is tailored to the IKL and keypoint estimation.

3. Incremental Object Keypoint Learning

3.1. Problem Formulation

We denote t=0 as the initial step and t=1, ..., T as the incremental steps. Then the training set for the t-th incremental step is $D_t^{\text{train}}=\{x_{t,j}^{\text{train}}, y_{t,j}^{\text{train}}\}_{j=1}^{N_t}$, where x, y, N_t denote the inputs, the keypoint labels and the size of D_t^{train}. As we only label new keypoints in IKL, thus the newly introduced keypoints must be disjointed with the old ones, i.e., $y_t^{\text{train}} \cap \left(\cup_{i=0}^{t-1} y_i^{\text{train}}\right) = \varnothing$. Different from the classification task, the common practice for keypoint estimation is to regress a 2D heatmap for a keypoint, i.e., a Gaussian peak around the ground-truth keypoint location [71]. Thus the model m_t in IKL comprises the feature extractor f_t and a stack of convolutional keypoint estimation heads $\{\mathrm{G}_i\}_{i=1}^{t}$ that parameterized by θ_t and $\{\phi_i\}_{i=1}^{t}$, respectively. Each estimation head ends with a convolutional layer to generate the 2D heatmap. G_t denotes the estimation head for the new keypoints in the t-th step. f_t and $\{\mathrm{G}_i\}_{i=1}^{t-1}$ in model m_t are initialized by the same weights in the old model m_{t-1}.

3.2. KAMP: A Novel Baseline for IKL

In this section, we present our two-stage method KAMP as a novel baseline tailored to IKL. We conjecture that distilling the relation between related old and new keypoints into the model may help it implicitly capture the intrinsic physical prior to alleviate the label non-co-occurrence issue. To do so, for each incremental step, in Stage-I, we design an auxiliary prediction task to associate the related

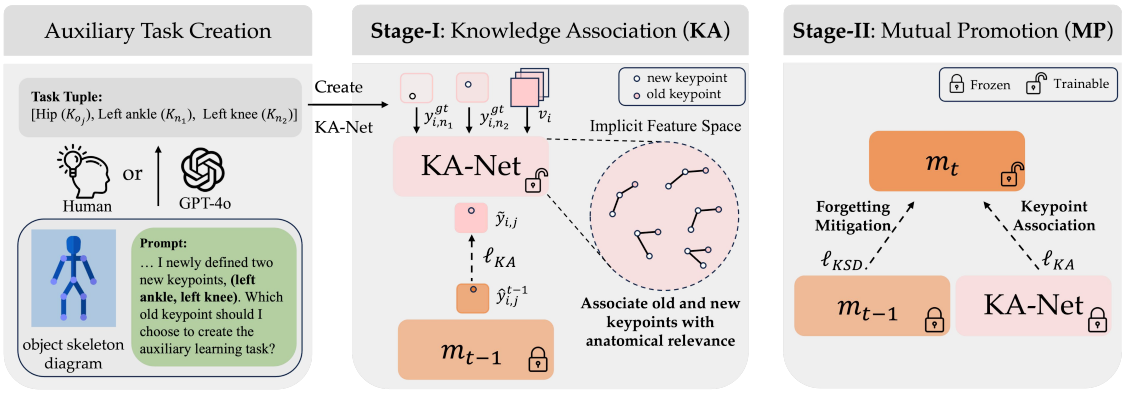

Figure 3. Overview of KAMP using the human body for illustration. In Stage-I, we learn an auxiliary KA-Net to associate the related old and new keypoints based on their local anatomical constraint. In Stage-II, we jointly leverage the old model and the KA-Net as an auxiliary teacher to consolidate all old keypoints' prediction and also learn the new keypoints simultaneously to achieve mutual promotions.

old keypoint's with the new ones by the KA-Net. In Stage-II, the old model m_{t-1} and KA-Net are both frozen to distill knowledge into the new model m_t for all old keypoints, and the new keypoints are learned concurrently to achieve mutual promotion. Fig. 3 provides a schematic view of KAMP.

3.2.1. Stage-I: Knowledge Association (KA)

As mentioned in Sec. 1, the old and new keypoints in IKL are not jointly labeled, making it difficult for the model to learn relationships between keypoints, such as their structural and anatomical relevance, using only the labels of new keypoints. To mitigate this issue, we propose to model the constraint among the spatially and anatomically related old and new keypoints as an implicit function in the current incremental step t. Existing analysises [11, 12, 64] show that keypoint association can be well-modeled by the triangulation constraint. Motivated by this, we define a triangulation constraint F on the distribution of three related keypoints:

$$F(P(K_i), P(K_j), P(K_k)) = 0, \quad (1)$$

where $P(K_i)$ denotes the distribution of the keypoint i, and the tuple may include one or two new keypoints. However, as there may be multiple valid solutions (constraints) that satisfy the implicit function F, it is challenging to derive an analytical solution for F in IKL.

Therefore, we consider an alternative formulation, e.g., $P(K_{o_j}) = F(P(K_{n_1}), P(K_{n_2}))$, which is a special case of the Eqn. 1 that we condition the distribution of an old keypoint $P(K_{o_j})$ on its related new keypoints $P(K_{n_1})$ and $P(K_{n_2})$. We focus on two new keypoints since only they have ground-truth labels, and the ground-truth supervision can help reduce the uncertainty in learning the constraints. Moreover, in practice, new tasks typically require more than one new keypoint for support. Nonetheless, we verify in our Supplementary that our approach performs effectively even with a single new keypoint. To enhance this constraint, we model the conditional probability between the related old and new keypoints as an auxiliary prediction task:

$$P(K_{o_j} \mid K_{n_1}, K_{n_2}, v) = \text{KA}(P(K_{n_1}), P(K_{n_2}), v), \quad (2)$$

where KA stands for the **K**nowledge **A**ssociation Network (KA-Net). The insight of this auxiliary task is inspired by the fact that spatially adjacent and visually correlated keypoints can mutually predict each other as shown in many existing works [11, 62, 64].

KA-Net. The input of the KA-Net is the ground-truth heatmap (y^{gt}_{i,n_1} and y^{gt}_{i,n_2}) of new keypoints K_{n_1} and K_{n_2} of an image i, and the output is the predicted heatmap $\tilde{y}_{i,j}$ of the old keypoint K_{o_j}. v_i denotes the holistic visual features v of the image i comprised of the intermediate features extracted from the frozen feature f_{t-1} and are re-scaled to the same spatial size as y^{gt}_{i,n_1} and y^{gt}_{i,n_2}. The visual features v_i are incorporated alongside spatial information (y^{gt}_{i,n_1} and y^{gt}_{i,n_2}) because, in a 2D image, spatial coordinates alone are insufficient to determine a keypoint's location due to uncertainty stemming from factors such as camera angle, object appearance, and motion. By incorporating visual features, we provide additional contextual information that helps account for this uncertainty in keypoint associations.

In the present paper, we explore KA-Net's simplest construction to minimize its training cost: for each new keypoint, we first perform the element-wise multiplication between its ground-truth heatmap (y^{gt}_{i,n_1} and y^{gt}_{i,n_2}) and the holistic features v_i to obtain the keypoint-oriented spatial features. Then we concatenate the spatial features and feed them forward over three convolutional layers accompanied by the Batch Normalization (BN) and ReLU to predict the selected old keypoint K_{o_j}. Note that the trained KA-Net will only be used to distill the keypoint association in the Stage-II and will **not** be used in test time after IKL.

Training of KA-Net. We train KA-Net by the new data D^{train}_t. As D^{train}_t does not have ground truth labels for old keypoints, we use the pseudo-labels predicted by the old model m_{t-1} to supervise KA-Net for keypoint regression:

$$\ell_{KA} = \frac{1}{N_t} \sum_i^{Nt} \sum_{j \in \mathcal{K}_{KA}} \left\| \hat{y}^{t-1}_{i,j} - \tilde{y}_{i,j} \right\|_2^2, \quad (3)$$

where $\mathcal{K}_{KA}$ denotes **the selected old keypoints** used for the auxiliary prediction task, $\hat{y}_{i,j}^{t-1}$ denotes the prediction of the j-th keypoint by m_{t-1} given image i, and $\tilde{y}_{i,j}$ denotes the prediction by the KA-Net. We will show in Sec. 4 that our KAMP can effectively reduce the forgetting of old keypoints and even improve them, which largely reduce the accumulative error to use pseudo-labels for KA-Net.

Auxiliary Task Creation. To select the old keypoint for training KA-Net, in each incremental step, we first locate all the old and new keypoints on a general object anatomy diagram by their semantic definition. Then we measure their relative proximity by simple distance calculations (e.g., Euclidean distance) to identify two new keypoints (K_{n_1} and K_{n_2}) that are close to an old one K_{o_j}. This process can also be automated by multi-modality large language models like GPT-4o [1, 42, 75] as shown in Fig. 3. If we identify several tuples of old and new keypoints satisfied the requirement, we randomly choose one tuple to create the auxiliary prediction task to avoid training too many KA-Net. For all results of our KAMP in Section 4, we show that even by creating only one auxiliary task in each step, we can still bring sufficient improvement on all keypoints. Note that task creation does not need any training, and we only need to perform once before each step, which is highly efficient. Such a design provides an interpretable way to incorporate the physical knowledge of the object to guide the IKL.

3.2.2. Stage-II: Mutual Promotion (MP)

In Stage-II, we jointly optimize all the new and old keypoints on the new model m_t by the loss l_{MP}:

$$l_{MP} = \ell_{GT} + \alpha\left(\ell_{KSD} + \ell_{KA}\right), \tag{4}$$

where ℓ_{GT} denotes the L2 loss between the ground truth of the new keypoints and their predictions by the new model to acquire the knowledge of new keypoints. α is a hyper-parameter to balance the new keypoints acquisition and old keypoints forgetting. For ℓ_{KA}, we use the frozen KA-Net as an auxiliary teacher to supervised the old keypoint selected in Stage-I to distill the keypoint association knowledge. Since ℓ_{KA} is only applied to the selected old keypoint for knowledge transfer instead of mitigating the forgetting, we further consolidate the knowledge of all old keypoints by the loss ℓ_{KSD} using the frozen old model m_{t-1}. As the distribution of new and old data in IKL do not change dramatically and may even have a strong correlation, LWF [40] can be a baseline for ℓ_{KSD} in such a scenario as analyzed in [13]. In general, ℓ_{KSD} represents the negative log-likelihood $\frac{1}{N_t}\sum_i^{N_t} s\left(\hat{y}_i^{t-1}\right) \cdot \log s\left(\hat{y}_i^{t}\right)$ between the predictions of all old keypoints by the new model, $\hat{y}_i^{t-1} \in \mathbb{R}^{C\times H\times W}$, and predictions by the old model, $\hat{y}_i^{t} \in \mathbb{R}^{C\times H\times W}$. $s(\cdot)$ denotes the Softmax operator. C denotes the numbers of old keypoints learned before step t, H and W are the height and width of each keypoint's heatmap.

However, since LWF and its variants all focus on the classification task, they perform the Softmax across different classes by default, i.e., over the C dimension, to obtain the normalized class prediction score. For IKL, this means normalizing across all old keypoints' predictions, which can not explicitly regularize the discrepancy of each old keypoint between the old and new model. To better preserve keypoint-specific knowledge, we adopt a spatial softmax operation over height (H) and width (W) dimensions of the keypoint prediction, $s_{\mathrm{sp}}^{\mathrm{d}}(\cdot), \mathrm{d} \in \{H, W\}$, and combine them to encourage spatial-oriented knowledge distillation. We discuss in our Supplementary the difference between each softmax alternative. Thus, our ℓ_{KSD} is defined as

$$\frac{1}{N_t}\sum_i^{Nt}\sum_j^{C}\sum_{\mathrm{d}} -s_{\mathrm{sp}}^{\mathrm{d}}\left(\hat{y}_{i,j}^{t-1}\right) \cdot \log s_{\mathrm{sp}}^{\mathrm{d}}\left(\hat{y}_{i,j}^{t}\right). \tag{5}$$

4. Experiment

Datasets. As the IKL has rarely been studied before, there is no specific dataset or benchmark for IKL. Motivated by the real-world application of IKL in medical analysis, we leverage the large-scale Cephalometric keypoint dataset [6] proposed in the 2023 MICCAI Challenge and also the commonly benchmarked Chest dataset [27] to further create the IKL protocols to validate our proposed method. Note that the Cephalometric keypoint dataset [6] differs from the head dataset [67] used in previous works [78, 79] commonly, as [6] has much larger number of keypoints and re-collected from different hospitals to enlarge the variance of images, makeing it much more challenging than [67]. Thus we term this dataset [6] as Head-2023 for clarity. Note that both Head-2023 and Chest datasets have less than or equal to 400 images in total. Moreover, though there is not a large need for IKL on human and animal, the widely used human and animal keypoint datasets, i.e., MPII [5] and ATRW [37], have extreme variations and non-rigid motions. Thus we also choose them for experiments to show the generality of our KAMP under different domains and challenging scenarios. We detail full dataset statistics in our Supplementary.

Compared Methods. For keypoint estimation, common network structures for classification like ResNet [21] need specific modifications, e.g., adding the deconvolution layers and convolution layers to generate the keypoint location heatmap. These adjustments make many IL methods inapplicable for IKL, e.g., prototype-based methods [87, 88] like PASS [88] that leverages the feature mean prototype before the linear classification layer. Furthermore, the IKL requires that the old data can not be retained, and it is hard to identify specific methods for IKL. Thus we choose several general and representative exemplar-free methods that can adapt to different IL settings based on their methodologies, including EWC [33], LWF [40], MAS [72], RW [9], AFEC [69], CPR [8] for comparison. They are regularization-based methods and

Method	Split Chest			Split Head-2023			Split MPII			Split ATRW		
	A-MRE$_1$ ↓	AT$_1$ ↑	MT$_1$ ↑	A-MRE$_4$ ↓	AT$_4$ ↑	MT$_4$ ↑	AAA$_4$ ↑	AT$_4$ ↑	MT$_4$ ↑	AAA$_3$ ↑	AT$_3$ ↑	MT$_3$ ↑
Joint Training	5.43	-	-	2.12	-	-	88.50	-	-	94.69	-	-
Finetune	43.1	-	-	51.3	-	-	37.41	-	-	13.24	-	-
EWC [33]	13.28	-8.23	-3.67	10.97	-6.37	-4.76	38.64	-51.84	-12.21	14.38	-59.75	-2.08
RW [9]	9.48	-7.12	-4.15	6.49	-4.23	-2.88	38.47	-18.83	-7.13	84.15	-10.87	0.00
MAS [4]	7.36	-1.86	-0.19	5.31	-2.15	-1.33	67.03	-7.56	0.34	85.68	-5.80	-1.13
LWF [40]	6.35	-1.34	0.18	4.31	-1.26	0.57	75.75	-3.86	0.41	87.31	-5.10	-0.64
AFEC [69]	8.04	-2.67	0.15	5.77	-3.45	-1.46	37.24	-22.85	-15.42	33.03	-40.25	-8.02
CPR [8]	6.17	-0.87	0.29	3.71	-1.18	0.16	75.52	-3.24	0.75	89.34	-2.76	4.49
SFD [17]	7.68	-0.54	0.13	4.76	-0.43	0.02	71.49	-0.93	0.21	86.11	-1.13	0.41
WF [74]	7.31	-0.31	0.16	4.58	0.03	0.11	72.87	-0.46	0.38	86.69	-0.97	0.62
GBD [16]	6.42	0.06	0.21	4.34	0.12	0.47	75.62	-0.18	0.35	87.42	-0.89	0.65
KAMP (Ours)	**5.67**	**0.29**	**0.62**	**2.32**	**0.41**	**0.84**	**79.93**	**1.80**	**4.23**	**93.16**	**-0.84**	**5.13**

Table 1. Result of 4 datasets after 2-Step, 5-Step, 5-Step and 4-Step IKL for Chest, Head-2023, MPII, and ATRW respectively. All comparison methods are started from the same Step-0 trained model. A-MRE: smaller the better; AAA/AT/MT: larger the better.

can be easily applied to the IKL without trial and error. We further adapt methods from Continual Semantic Segmentation (CSS) and Class Incremental Learning (CIL) to ISL for more comparisons, i.e., spatial feature distillation (SFD) [17], weight fusion (WF) [74], and gradient balanced distillation (GBD) [16]. We also report the native baseline, i.e., directly finetune the model during IKL, and the upper bound that trains the model (e.g., HRNet [60]) with all the data, where we denote the former as "Finetune" and the latter as "Joint Training".

Evaluation Metrics. To assess the keypoint regression task, we employ the widely used mean radial error (MRE) for Head-2023 and Chest datasets as in [78, 79], and Probability of Correct Keypoint (PCK) [50, 56, 66, 77] for MPII and ATRW and use their defaulted σ as in [50, 60] to compute PCK. To measure the performance of incremental learning (IL), we use Average Accuracy (AAA), calculating the accuracy (%) across all keypoints post-incremental step i under PCK, and extend this approach to MRE, denoting it as A-MRE. Additionally, we introduce two metrics for knowledge transfer in IL: (1) **Average Transfer (AT)**, also known as backward transfer [47], which averages the performance improvement of keypoints over all previous steps after step i, and (2) **Maximal Transfer (MT)**, measuring the largest performance improvement in any old keypoint post-step i. Notably, when calculating AT and MT with MRE, we invert the sign of the change in error. This adjustment ensures consistency, since a decrease in MRE signifies improvement, in contrast to the direct correlation of increased accuracy with improvement in the PCK metric. Detailed explanations for the calculation of each metric are provided in our supplementary materials.

Experimental Design. We randomly split the keypoints of Chest, Head-2023, MPII, and ATRW into different incremental steps to create the Spilt Head-2023, Chest, MPII and ATRW protocols respectively. As the Chest dataset only has 6 keypoints in total, we split them into two groups where the first group has 4 keypoints with 150 training images, and the second group for IKL has 2 keypoints with only 50 training images. For Split Head-2023, we split the 38 keypoints into 5 groups, where the first group has 19 keypoints with the same definition as [67] with 100 training images, and we split the rest of keypoint into 4 groups randomly to simulate four incremental steps with only 50 training images for each step. Similarly, the 16 MPII keypoints and 15 ATRW keypoints were split into 5 and 4 groups, respectively, with an equal distribution of training images per step. We repeat five times under different orders of keypoint groups and report the mean value in the main paper. The standard deviation (std), details about the keypoints group, and results of other experimental setups are in our Supplementary.

Implementation Details. We use the HRNet [60] as the backbone for all methods as it is widely compared in keypoint estimation [19, 36, 39, 50, 66, 79]. For a fair comparison, we initialize all methods with the same initial step (Step-0) model for incremental learning. As the first benchmark for the IKL, we use the Continual Hyperparameter Framework (CHF) [13, 41] to identify the training parameters like training epoch (100), initial learning rate (2e-3 for Split Chest and Head-2023, 1e-2 for Split MPII, 1e-3 for Split-ATRW), momentum (0.9), weight decay (1e-4), and also the hyperparameters of each compared methods, all included in our Supplementary. For our KAMP, the hyperparameter α is set as 1e2 for the Split Head-2023 and Split-MPII and 1e4 for the Split Chest and Split-ATRW in all experiments. The analysis of the α and other implementation details are included in our Supplementary.

4.1. Comparison with SOTA for IKL

As shown in Tab. 1, our KAMP consistently improves accuracy and reduces error across all old and new keypoints after each incremental step, achieving the highest performance compared to all other methods on all datasets. For example, KAMP outperforms the second-best method by **1.39%** in A-MRE$_4$ after four incremental steps on Split Head-2023, and by **4.18%** in AAA$_4$ on Split MPII. Moreover, KAMP is the only method to consistently yield the highest average transfer (AT) and maximal transfer (MT) scores across all datasets. This result highlights that our two-stage learning scheme better facilitates knowledge transfer, thereby en-

hancing accuracy on old keypoints while learning new ones. When AT is positive, this indicates *no forgetting on average*, underscoring the effectiveness of using pseudo-labels from the old model to train KA-Net, as described in Sec. 3.2.1. In Split-ATRW, the AT is negative for all methods, indicating that forgetting of old keypoints outweighs knowledge transfer. However, KAMP still achieves the highest AT among the methods. Notably, the absolute value of KAMP's AT is nearly zero, suggesting minimal negative transfer. Additionally, by examining the maximal transfer metric, we observe a significant positive transfer in specific cases, such as **5.13%** for MT_3, indicating that some old keypoints experience a substantial positive transfer in each incremental step. This positive transfer helps offset the forgetting of other old keypoints, resulting in a minimal negative average transfer. We further validate this effect by reporting per-keypoint performance in our Supplementary. Qualitative results in Fig. 4 further verify our superiority in achieving more structurally correct prediction than competitive methods.

Lastly, we observe that methods adapted from CSS and CIL, i.e., SFD, WF, and GBD, achieve strong performance on AT and MT metrics compared to competitive methods like LWF and CPR, but perform poorly on overall metrics like A-MRE and AAA. This is because these methods prioritize reducing forgetting, resulting in an overly rigid model that struggles to effectively incorporate new keypoints in IKL. This observation highlights a crucial limitation of existing CSS and CIL methods that they focus heavily on regularizing forgetting but fail to balance this with knowledge acquisition for new keypoints, which is especially problematic in our newly defined IKL setting. Consequently, this emphasizes the need to study IKL as a distinct incremental learning scenario in this paper, as traditional incremental learning methods are too general to excel here. **Our proposed novel baseline, KAMP, effectively addresses this gap for IKL, offering a well-balanced approach that goes beyond anti-forgetting to deliver robust performance across both accuracy and transfer metrics.**

4.2. Ablation Study

To verify the effectiveness of our method, we further analyze how each component may influence the result. Since our method contains two-stage training and we consider a keypoint-oriented spatial distillation loss (ℓ_{KSD}), thus we compared our method with three alternatives: (1) the competitive baseline, i.e., the LWF, which performs the Softmax operation across old keypoints. (2) only use our ℓ_{KSD} to train the model without the KA-Net. (3) Constructing the KA-Net **randomly** without using physical knowledge. Results are shown in Tab. 2, and we can observe that: (1) our adapted ℓ_{KSD} can effectively outperform the LWF with 1.18%, which demonstrates the essence of our keypoint-oriented adaptation. (2) With proper auxiliary task creation based on physical prior, we can achieve more positive average transfer than other counterparts and finally achieve the largest improvement over all the keypoints on AAA_4. This shows that our proposed two-stage learning scheme for IKL can not just provide better knowledge consolidation on the old keypoints than the competitive baseline, but the auxiliary prediction task can also bring large improvement to the old keypoints, showing that our KAMP will be a novel and strong baseline for the IKL paradigm. More ablation studies like using backbones other than HRNet and more datasets are in our Supplementary.

Method	$AAA_4 \uparrow$	$AT_4 \uparrow$	$MT_4 \uparrow$
LWF [40]	75.75	-3.86	0.41
KAMP (only ℓ_{KSD})	76.93	-2.24	0.65
KAMP (Random KA-Net)	77.13	-0.48	1.24
KAMP (Ours)	**79.93**	**1.80**	**4.23**

Table 2. Ablation Study on Split MPII.

4.3. Compare IKL to other low-data paradigms

Now we provide more insights into our IKL by comparing it with other learning paradigms with low-shot data. Discussion of limitations are in our Supplementary.

As mentioned before, existing learning paradigms like UKL [22, 23, 78] and CAPE [10, 56, 76] leverage large-scale self-supervised and multi-dataset training to enable novel keypoint detection with a few labeled samples in test time. To fairly compare with UKL and CAPE, we extend KAMP to low-data regime following EGT [78], where we also pretrain an auxiliary self-supervised model at Step-0 to provide pseudo-labels for old keypoints during IKL, and the details are in the Supplementary. For medical datasets like Head-2023, we compare to the SOTA one-shot methods CC2D [78] and EGT [79]. For 2D datasets like MPII, we compare to the SOTA Autolink [22] and MetaPoint+ [10] with official implementation available. Note that methods like [86] require extra expensive 3D CAD models to identify keypoint proposals for novel keypoint detection and are inferior to the SOTA [10, 56], thus we only compare to the MetaPoint+ [10]. For KAMP, we use the low-data adaptation for 1, 5, and 10 shot and do not use it for 50-shot since 50-shot is large enough for us to bypass the adaptation. Full experimental details are in our Supplementary.

Method/(MRE ↓)	1-shot	5-shot	10-shot	50-shot
CC2D [78]	5.14	4.83	4.08	3.47
EGT [79]	5.01	4.58	3.87	3.21
KAMP (Ours)	**4.35**	**3.70**	**3.03**	**2.32**

Table 3. Compare to low-shot methods on Split Head-2023.

From Tab. 3 we observe that even on the extreme 1-shot setting, KAMP still outperforms the CC2D and EGT, while

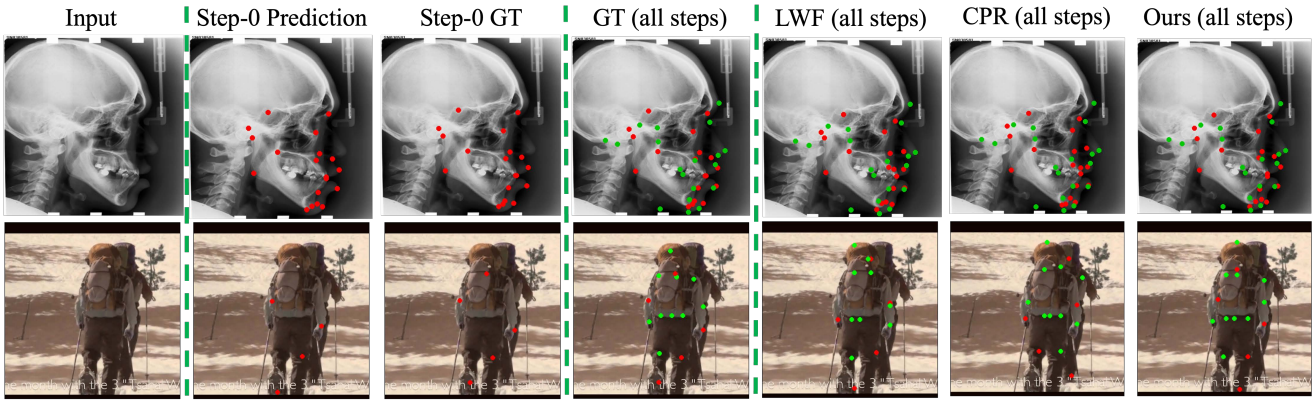

Figure 4. Qualitative results on Split Head-2023 and MPII. All methods start from the same Step-0 model, whose prediction is in the second column. GT: ground truth. The red circles denote the keypoints learned in Step 0, and the green ones denote all the new keypoints learned in later incremental steps. We observe that after 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. More results are included in our Supplementary.

all the compared methods do not have good performances. This is because the Head-2023 dataset is collected from multiple sources with large discrepancy, and only one image is hard to represent all the variations of the data. However, when more annotations are available, KAMP can scale much better than CC2D and EGT since our two-stage learning scheme can capture the relation of old and new keypoints and distill them back to the model to improve the old keypoints, which has not been handled in CC2D and EGT.

Method/(PCK [10] ↑)	1-shot	5-shot	10-shot	50-shot
UKL [22]	54.95	59.11	62.53	64.36
MetaPoint+ [10]	65.71	66.87	67.26	68.98
KAMP (Ours)	**70.09**	**72.23**	**73.18**	**76.97**
KAMP (Ours) + [10]	**73.49**	**75.10**	**77.76**	**79.18**

Table 4. Compare to low-shot methods on Split MPII.

As shown in Tab. 4, we observe that while the CAPE [10, 56, 76] and IKL are two disentangled paradigms, they can complement each other effectively. CAPE relies on pretraining with varied datasets, while IKL focuses on incremental learning, making their training costs incomparable. For example, CAPE methods like MetaPoint+ [10] require a substantial training cost (4 GPUs, 1.5 days), whereas the 'pretrain+incremental' approach of IKL is more efficient (1 GPU, 5 hours). To demonstrate that CAPE and IKL are orthogonal and complementary, we applied IKL using our KAMP method on MetaPoint+. As shown in Tab. 4, this combined approach results in significant improvements, even in low-data scenarios, highlighting that IKL can substantially enhance keypoint estimation when applied on a CAPE-pretrained keypoint detector.

The goal of comparing these paradigms for novel keypoint estimation within the same object category is to establish the necessity of IKL. Although CAPE does not involve continual learning and appears more deploy-friendly, its performance plateaus and does not scale effectively with additional data. In contrast, our KAMP achieves higher accuracy with limited data and demonstrates superior scalability with increased data, as evidenced in Tab. 4.

Lastly, as the first study to explore the IKL paradigm, we focus on scaling up existing keypoint estimators for novel keypoints within the same object category, laying the foundation for future research. While CAPE can perform keypoint estimation on novel object categories with a few support images, its generalization remains limited. In the future, we will explore extending the IKL paradigm to novel object categories by applying IKL to keypoint estimators pretrained with CAPE, aiming to combine the strengths of both paradigms to develop a keypoint estimator capable of continually learning new keypoints across diverse object categories without forgetting and with effective scalability.

In summary, the comparison of Tab. 3 and 4 show that our proposed IKL paradigm is highly label-efficient for acquiring new keypoints. This characteristic is advantageous for real-world applications where obtaining sufficient labels is time-consuming, such as in medical analysis [6].

5. Conclusion

We explore learning the newly defined keypoint incrementally without retaining any old data, called Incremental object Keypoint Learning (IKL). We propose a two-stage learning scheme as a novel baseline tailored to the IKL. Extensive experiments show that our method can effectively alleviate the forgetting issue and may even improve the old keypoints' estimation during IKL. Our further analysis reveals that the IKL is label efficient in acquiring the new keypoints, which is promising for real-world applications.

Acknowledgements This work was supported in part by National Science Foundation grant IIS-2007613. This work was also supported by the National Natural Science Foundation of China (62376011).

References

[1] Hello gpt-4o, howpublished = https://openai.com/index/hello-gpt-4o/. 5

[2] Hongjoon Ahn, Sungmin Cha, Donggyu Lee, and Taesup Moon. Uncertainty-based continual learning with adaptive regularization. In *Advances in Neural Information Processing Systems*, pages 4392–4402, 2019. 3

[3] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[4] 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. 6

[5] 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. 1, 2, 3, 5

[6] Jun Cao, Juan Dai, Xuguang Li, Bingsheng Huang, Ching-Wei Wang, and Hongyuan Zhang. Cephalometric landmark detection in lateral x-ray images, 2023. 1, 2, 5, 8

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[8] 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. 5, 6

[9] 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. 5, 6

[10] 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. 3, 7, 8

[11] Xiao Chu, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Structured feature learning for pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4715–4723, 2016. 4

[12] Shengyang Dai, Ming Yang, Ying Wu, and Aggelos Katsaggelos. Detector ensemble. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1, 4

[13] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3, 5, 6

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 3

[15] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5138–5146, 2019. 3

[16] 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. 6

[17] 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. 6

[18] Ana Paula Reis Durão, Aline Morosolli, Pisha Pittayapat, Napat Bolstad, Afonso P Ferreira, and Reinhilde Jacobs. Cephalometric landmark variability among orthodontists and dentomaxillofacial radiologists: a comparative study. *Imaging science in dentistry*, 45(4):213–220, 2015. 1

[19] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14676–14686, 2021. 1, 6

[20] 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. 3

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[22] 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. 1, 7, 8

[23] Xingzhe He, Gaurav Bharaj, David Ferman, Helge Rhodin, and Pablo Garrido. Few-shot geometry-aware keypoint localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21337–21348, 2023. 1, 7

[24] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2018. 3

[25] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via

rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019. 3

[26] Zhongzhan Huang, Mingfu Liang, Senwei Liang, and Shanshan Zhong. Flat local minima for continual learning on semantic segmentation. In *International Conference on Multimedia Modeling*, pages 388–400. Springer, 2025. 3

[27] 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. 2, 5

[28] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. *arXiv preprint arXiv:2207.13784*, 2022. 1

[29] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, pages 196–214. Springer, 2020. 1, 3

[30] Ta-Chu Kao, Kristopher Jensen, Gido van de Ven, Alberto Bernacchia, and Guillaume Hennequin. Natural continual learning: success is a journey, not (just) a destination. *Advances in Neural Information Processing Systems*, 34: 28067–28079, 2021. 3

[31] Leonid Karlinsky and Shimon Ullman. Using linking features in learning non-parametric part models. In *European Conference on Computer Vision*, pages 326–339. Springer, 2012. 1

[32] Chris Dongjoo Kim, Jinseo Jeong, and Gunhee Kim. Imbalanced continual learning with partitioning reservoir sampling. In *European Conference on Computer Vision*, pages 411–428. Springer, 2020. 3

[33] 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. 3, 5, 6

[34] 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. 3

[35] 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. 3

[36] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1944–1953, 2021. 1, 6

[37] 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. 2, 5

[38] 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. 3

[39] 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. 1, 3, 6

[40] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 3, 5, 6, 7

[41] Mingfu Liang, Jiahuan Zhou, Wei Wei, and Ying Wu. Balancing between forgetting and acquisition in incremental subpopulation learning. In *European Conference on Computer Vision*, pages 364–380. Springer, 2022. 3, 6

[42] Mingfu Liang, Jong-Chyi Su, Samuel Schulter, Sparsh Garg, Shiyu Zhao, Ying Wu, and Manmohan Chandraker. Aide: An automatic data engine for object detection in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14695–14706, 2024. 5

[43] Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Beyond not-forgetting: Continual learning with backward knowledge transfer. In *Advances in Neural Information Processing Systems*, 2022. 3

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 3

[45] Buyu Liu and Vittorio Ferrari. Active learning for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4363–4372, 2017. 1

[46] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26. PMLR, 2017. 3

[47] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pages 6467–6476, 2017. 3, 6

[48] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13264–13273, 2021. 3

[49] 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. 1, 3

[50] Olga Moskvyak, Frederic Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Semi-supervised keypoint localization. In *International Conference on Learning Representations*, 2021. 1, 3, 6

[51] Gaurav Kumar Nayak, Het Shah, and Anirban Chakraborty. Incremental learning for animal pose estimation using rbf k-dpp. *arXiv preprint arXiv:2110.13598*, 2021. 3

[52] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020. 3

[53] Ricketts Robert. *Orthodontic Diagnosis and Planning : – Their Roles in Preventive and Rehabilitative Dentistry*. Denver Colo; Rocky Mountain/Orthodontics, 1982. 1

[54] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 397–403, 2013. 3

[55] Enrique Sanchez and Georgios Tzimiropoulos. Object landmark discovery through unsupervised adaptation. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[56] 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. 3, 6, 7, 8

[57] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416, 2018. 1

[58] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 1

[59] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2

[60] 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. 6

[61] Wei Tang and Ying Wu. Does learning specific features for related parts help human pose estimation? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1107–1116, 2019. 3

[62] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 190–206, 2018. 3, 4

[63] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE international conference on computer vision*, pages 5916–5925, 2017. 3

[64] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27, 2014. 4

[65] Riccardo Volpi, Diane Larlus, and Grégory Rogez. Continual adaptation of visual representations via domain randomization and meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4443–4453, 2021. 3

[66] 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. 3, 6

[67] Ching-Wei Wang, Cheng-Ta Huang, Jia-Hong Lee, Chung-Hsing Li, Sheng-Wei Chang, Ming-Jhih Siao, Tat-Ming Lai, Bulat Ibragimov, Tomaž Vrtovec, Olaf Ronneberger, Philipp Fischer, Tim F Cootes, and Claudia Lindner. A benchmark for comparison of dental radiography analysis algorithms. *Med. Image Anal.*, 31:63–76, 2016. 1, 5, 6

[68] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1, 3

[69] 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. 5, 6

[70] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[71] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 3

[72] Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. In *Advances in Neural Information Processing Systems*, pages 5962–5972, 2018. 5

[73] Guile Wu, Shaogang Gong, and Pan Li. Striking a balance between stability and plasticity for class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1124–1133, 2021. 3

[74] 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 Vision and Pattern Recognition*, pages 7204–7213, 2023. 6

[75] Xiaoying Xing, Mingfu Liang, and Ying Wu. Toa: Task-oriented active vqa. *Advances in Neural Information Processing Systems*, 36:54061–54074, 2023. 5

[76] 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. 2, 3, 7, 8

[77] 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. 1, 3, 6

[78] 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. 3, 5, 6, 7

[79] 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. 5, 6, 7

[80] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. AP-10k: A benchmark for animal pose estimation in the wild. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 1, 3

[81] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6982–6991, 2020. 3

[82] Duncan Zauss, Sven Kreiss, and Alexandre Alahi. Keypoint communities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11057–11066, 2021. 3

[83] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. *Proceedings of Machine Learning Research*, 70:3987, 2017. 3

[84] 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. 1, 3

[85] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13208–13217, 2020. 3

[86] Xingyi Zhou, Arjun Karpur, Linjie Luo, and Qixing Huang. Starmap for category-agnostic keypoint and viewpoint estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–334, 2018. 7

[87] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-lin Liu. Class-incremental learning via dual augmentation. *Advances in Neural Information Processing Systems*, 34:14306–14318, 2021. 5

[88] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021. 5

[89] Xu Zou, Sheng Zhong, Luxin Yan, Xiangyun Zhao, Jiahuan Zhou, and Ying Wu. Learning robust facial landmark detection via hierarchical structured ensemble. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 141–150, 2019. 3