PoseAdapt: Sustainable Human Pose Estimation via Continual Learning Benchmarks and Toolkit

Muhammad Saif Ullah Khan and Didier Stricker
German Research Center for Artificial Intelligence (DFKI)
Trippstadter Str. 122, 67663 Kaiserslautern, DE
https://saifkhichi96.github.io/research/poseadapt/

Human pose estimators are typically retrained from scratch or naively fine-tuned whenever keypoint sets, sensing modalities, or deployment domains change—an inefficient, compute-intensive practice that rarely matches field constraints. We present **PoseAdapt**, an open-source framework and benchmark suite for *continual pose model adaptation*. PoseAdapt defines domain-incremental and class-incremental tracks that simulate realistic changes in density, lighting, and sensing modality, as well as skeleton growth. The toolkit supports two workflows: (i) *Strategy Benchmarking*, which lets researchers implement continual learning (CL) methods as plugins and evaluate them under standardized protocols; and (ii) *Model Adaptation*, which allows practitioners to adapt strong pretrained models to new tasks with minimal supervision. We evaluate representative regularization-based methods in single-step and sequential settings. Benchmarks enforce a fixed lightweight backbone, no access to past data, and tight per-step budgets. This isolates adaptation strategy effects, highlighting the difficulty of maintaining accuracy under strict resource limits. PoseAdapt connects modern CL techniques with practical pose estimation needs, enabling adaptable models that improve over time without repeated full retraining.

1. Introduction

Human pose estimation enables applications across autonomous systems [16, 60, 67], healthcare [10, 64, 68], sports analytics [20, 63], and human–computer interaction [3, 19, 55]. Advances in representation methods [29, 47, 57, 66], model architectures [21, 37, 62], data scale [1, 22, 25, 32, 44], and training pipelines [23, 40] have pushed 2D keypoint benchmarks toward saturation [14].

Despite these gains, state-of-the-art models are fundamentally static: they are trained once on fixed datasets and deployed under the assumption that test distributions match training. In practice, changes in illumination, viewpoint, density, or sensing modality cause significant drops in accuracy [8, 12]. These limitations are especially visible in dynamic or resource-constrained settings (e.g., high-speed sports [20, 63] or egocentric capture [18, 59]), where strong motion, occlusion, and field-of-view shifts challenge standard training regimes. These domains often involve domain-specific skeletons or sensing modalities for which

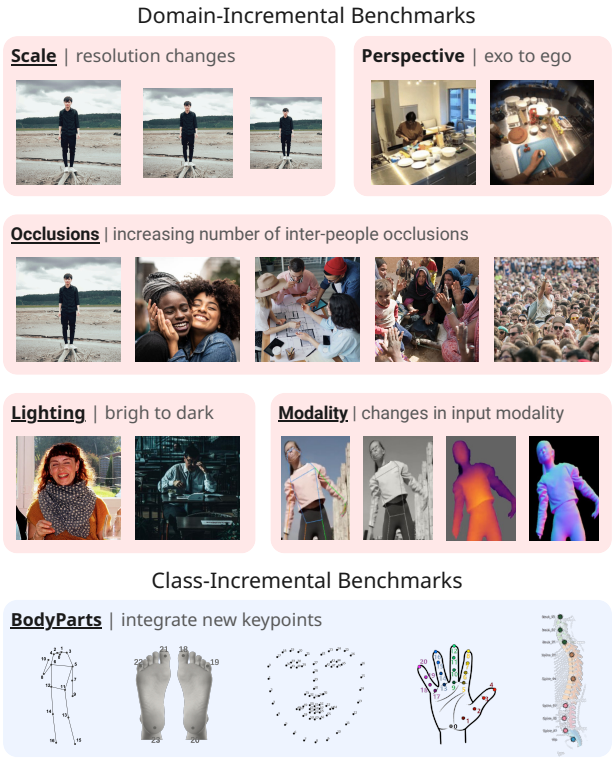


Figure 1. **PoseAdapt Benchmarks.** We introduce a diverse suite of domain- and class-incremental benchmarks for human pose estimation. *Top*: Domain-incremental settings simulate increasing difficulty through scale, perspective, occlusion, lighting, and modality shifts. *Bottom*: Class-incremental benchmarks gradually add new keypoint types to evaluate the ability to extend skeletons over time. All benchmarks share a fixed backbone, fixed per-step data budget, and unified evaluation protocol.

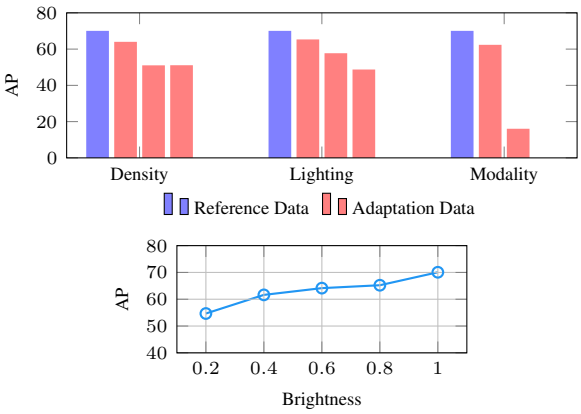


Figure 2. **Off-the-shelf models struggle under realistic shifts.** *Top:* Accuracy on the pretrained reference dataset (blue) drops consistently under sequential shifts in density, lighting, and modality (red), even though these changes are relatively minor compared to training conditions. *Bottom*: When brightness is progressively reduced on the same dataset, AP declines steadily, underscoring the brittleness of static models to illumination variation.

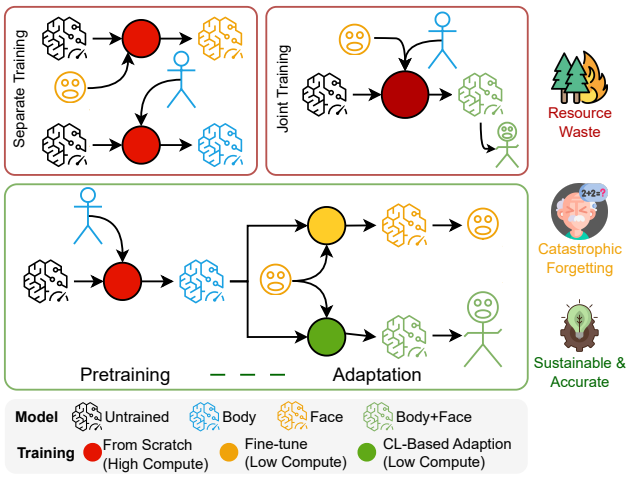


Figure 3. **Adaptation strategy comparison** *Top*: Conventional solutions either train models separately (resource waste) or fine-tune (prone to forgetting). *Bottom*: PoseAdapt enables adaptation using continual learning (CL) techniques to retain prior knowledge while specializing to new skeletons or domains.

pretrained estimators are either mismatched or incomplete. As shown in Fig. 2, off-the-shelf models degrade under realistic shifts, undermining their deployment-time performance despite high accuracy on benchmark datasets.

To overcome these deployment-time limitations or skeleton mismatches, off-the-shelf models often need to be fine-tuned on domain-specific data before they can be integrated in a real-world application. In Fig. 3, we compare different adaptation strategies used in practice. Retraining a model from scratch for each deployment condition is costly and slow, and naive fine-tuning tends to overwrite prior knowledge, leading to catastrophic forgetting [26, 30]. Some recent efforts target cross-skeleton generalization [9], but typically rely on large backbones or extensive supervision, limiting deployability on edge devices. In contrast, continual learning (CL) methods regularize updates to preserve previously acquired knowledge while specializing to new experiences [26, 30]. Therefore, *we advocate continual adaptation as a sustainable alternative.* Instead of discarding prior competence, models should incrementally incorporate new domains or keypoints while retaining past performance.

To this end, **we propose PoseAdapt, which operationalizes the continual adaptation principle for human pose estimation**. This includes (1) an open-source framework for adapting pretrained estimators to domain-specific datasets using established CL strategies, and (2) a dedicated benchmark for comparing performance of different CL strategies under strict, deployment-oriented constraints for both domain-incremental and class-incremental scenarios. Our PoseAdapt framework allows practitioners to trivially fine-tune off-the-shelf estimators on their application-specific data in a sustainable manner. Concurrently, the principled PoseAdapt benchmark encourages researchers to develop novel CL strategies for human pose estimation tasks by enabling fair and reproducible comparison.

The plug-and-play design of our PoseAdapt framework supports future extensions to new adaptation methods. This provides an evolving resource for human pose estimation practitioners and researchers which formalizes continual learning for human pose estimation.

Contributions.

1. We introduce PoseAdapt, an open-source framework for continual learning in pose estimation, with support for domain- and class-incremental scenarios[1].
2. We design challenging, realistic benchmark protocols that capture gradual distribution shifts in resolution, occlusion, lighting, modality, and skeleton structure.
3. We release modular toolkit with dataset wrappers, plugin-based CL strategies, and protocol-aware evaluation tools to facilitate sustainable pose model adaptation research.

2. Related Work

Pose Estimation Toolkits and Distribution Shifts. Open-source systems such as OpenPose [5], Detectron2 [61], MMPose [43], AlphaPose [13], RTMPose [21], and ViTPose [62] have driven rapid progress in 2D human pose estimation with modular training and evaluation across datasets and backbones. Despite their breadth, these toolk-

[1]Source code available at https://github.com/dfki-av/poseadapt/. Experiments reported in this paper use code from the "wacv-2026-camera-ready" branch.

its largely assume a static train–test regime in which models are trained once on curated data and deployed unchanged. Related adaptation efforts in vision—for example, test-time adaptation via entropy minimization [58]—offer robustness within a domain but do not address *incremental* task streams, skeleton growth, or long-horizon retention. Some cross-skeleton methods jointly train on multiple datasets [51], which is inefficient and impractical for continual evolution. Consequently, deploying pose estimators in dynamic settings (wearables, robotics, sports analytics) remains brittle when illumination, viewpoint, density, or sensing modality shift over time. *PoseAdapt framework provides continual adaptation protocols for pose estimation, missing in existing systems.*

Benchmarks and Datasets for Pose Estimation. Large-scale benchmarks (COCO Keypoints [31], MPII [1]) and newer efforts (COCO-WholeBody [22], PoseTrack [2], Halpe [13], SpineTrack [25]) have standardized evaluation and spurred architectural advances. However, these benchmarks are organized around static splits and assume full retraining when the task changes (e.g., skeleton extensions or cross-modality transfer). In contrast, continual learning (CL) benchmarks in classification study nonstationary streams and forgetting using protocols such as Split-CIFAR [65], CORe50 [33], and DomainNet [46]. *An analogous, pose-specific benchmark that enforces realistic constraints (fixed lightweight backbones, no access to past data, limited budgets) has been lacking; PoseAdapt fills this gap with domain- and class-incremental tracks.*

Continual Learning: Paradigms and Constraints. CL aims to learn from nonstationary streams without catastrophic forgetting [41, 45, 56]. Representative approaches include: *regularization-based* methods that constrain parameter or function drift (EWC [26], SI [65], LwF [30], LFL [24], IMM [28]); *replay-based* methods that use exemplars or generative rehearsal (iCaRL [49], DGR [53], A-GEM [7], DER [4]); and *architectural* approaches that isolate or grow capacity (PNNs [50], PackNet [38], Piggyback [39], HAT [52]). Many of these assume either access to a replay buffer or the ability to expand parameters and heads per task, which conflicts with deployment constraints such as fixed memory, strict compute, and privacy limits on past data. *PoseAdapt benchmark isolates adaptation strategy effects by fixing model architecture, disallowing access to old data, and enforcing per-step budgets, enabling fair comparison under realistic constraints.*

Continual Learning for Dense Prediction. Transferring CL to dense prediction introduces unique challenges: spatial outputs, structured losses, and label-evolution (e.g., new classes or keypoints). Recent work in segmentation [6, 42] and detection [54] adapts regularization, distillation, or pseudo-labeling to retain prior categories while learning new ones. For human pose estimation, CL remains comparatively underexplored; early studies adapt CL ideas to structured outputs [15], but evaluations are heterogeneous, often using small-scale or ad hoc protocols. *By standardizing domain-incremental and class-incremental settings with shared budgets and metrics (forward transfer, retention), PoseAdapt provides the first controlled testbed to assess CL strategies for pose at scale.*

Beyond benchmarks, software ecosystems for CL such as Avalanche [34] and Continuum [11] provide abstractions and baselines, while general domain-shift suites (e.g., WILDS [27]) and domain generalization toolkits (e.g., DomainBed [17]) target classification or detection tasks. MMPose and Detectron2 offer strong pose tooling but assume static training. *PoseAdapt complements these ecosystems by bringing pose-specific continual protocols, plug-in CL strategies, and unified evaluation under deployment-motivated constraints, enabling reproducible comparisons and practical adaptation pathways.*

3. PoseAdapt Framework

PoseAdapt is designed as a general-purpose continual adaptation layer on top of MMPose, enabling pretrained estimators to evolve across a sequence of experiences without requiring modifications to underlying backbones or datasets. It provides a unified mechanism for expressing how a model should initialize, update, and regularize itself as new experiences arrive. The goal is to decouple the mechanics of continual adaptation from the specifics of any particular backbone, dataset, or training regime.

Problem Setting. We consider a stream of experiences $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_T$, where each experience $\mathcal{E}_i$ provides a dataset $\mathcal{D}_i$ drawn either from a new domain (e.g. lighting, modality) or an expanded keypoint set. A pretrained off-the-shelf pose estimator $\mathcal{M}_0$ serves as the starting point. At each stage, the framework produces an updated model $\mathcal{M}_i = \Phi(\mathcal{M}_{i-1}, \mathcal{D}_i; \pi)$, where π denotes the chosen continual learning strategy. This formulation explicitly separates three roles: the backbone architecture contained in $\mathcal{M}_{i-1}$, the new information encoded in $\mathcal{D}_i$, and the adaptation policy encoded in π. Each experience is decomposed into three stages: (1) *Initialization*: prepare the structure and reference state required by the strategy; (2) *Adaptation*: optimize parameters on $\mathcal{D}_i$ under constraints of π; and (3) *Finalization*: compute strategy-specific statistics used by the next experience. Figure 4 illustrates this cycle: a new experience triggers structural initialization and then constrained adaptation. The right side shows how head expansion preserves output-space compatibility during class-incremental updates.

3.1. Initialization Phase

Given $\mathcal{E}_i$, PoseAdapt prepares $\mathcal{M}_{i-1}$ for the new experience.

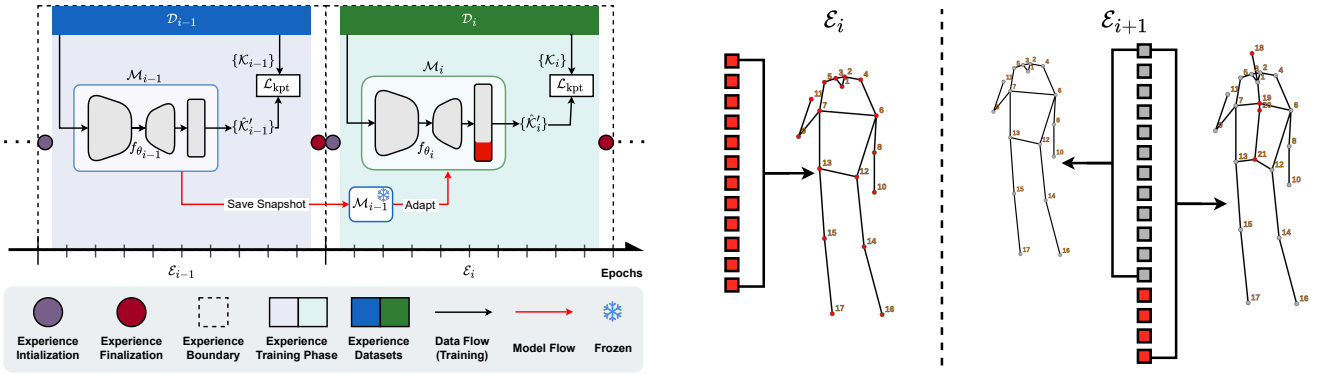


Figure 4. **PoseAdapt Framework.** *Left:* At each experience $\mathcal{E}_i$, PoseAdapt initializes the model for the new experience (e.g., snapshot creation, head expansion, or architecture-specific adjustments), followed by an adaptation phase that optimizes the model on $\mathcal{D}_i$ with strategy-defined regularization, and a finalization step to compute and store any statistics for later experiences. *Right:* Head expansion for class-incremental experiences, where the output dimensionality grows as new keypoints are introduced.

For fixed-architecture strategies (LwF, LFL, EWC), a frozen reference snapshot $\tilde{\mathcal{M}}_{i-1}$ is created. This snapshot provides target features (LFL), target logits (LwF), or parameter anchors (EWC).

In class-incremental settings, the prediction head expands from K_{i-1} to K_i:

$$W_i = \left[W_{i-1} \;\; \Delta W_i \right],$$

with ΔW_i initialized via configurable policy.

3.2. Adaptation Phase

Let θ_i denote the trainable parameters during experience i. The supervised loss on $\mathcal{D}_i$ is $\mathcal{L}_{\text{kpt}}(\theta_i; \mathcal{D}_i)$. Continual learning adds a strategy-dependent regularizer:

$$\mathcal{L}(\theta_i) = (1-\alpha)\,\mathcal{L}_{\text{kpt}} + \alpha\,\mathcal{L}_{\text{reg}}(\theta_i; \tilde{\mathcal{M}}_{i-1}; \pi).$$

Less-Forgetful Learning (LFL). LFL constrains the feature extractor to preserve the geometry learned previously. Let $f_i(x)$ and $\tilde{f}_{i-1}(x)$ denote backbone feature maps from the current and teacher models. The penalty is the per-level MSE:

$$\mathcal{L}_{\text{reg}}^{\text{LFL}} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \sum_{\ell} \left\| f_i^{(\ell)}(x) - \tilde{f}_{i-1}^{(\ell)}(x) \right\|_2^2.$$

Learning without Forgetting (LwF). LwF distills the teacher's output behavior. Let $z_i(x)$ and $\tilde{z}_{i-1}(x)$ denote logits from the current and teacher models (aligned by an optional converter if keypoints change). PoseAdapt applies softened KL divergence with confidence masking:

$$\mathcal{L}_{\text{reg}}^{\text{LwF}} = \tau^2 \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \text{KL}(\sigma(z_i(x)/\tau) \parallel \sigma(\tilde{z}_{i-1}(x)/\tau)).$$

EWC (online). EWC penalizes deviation from previous parameter values based on Fisher importances. Let θ_{i-1} be parameters after the last experience, and let $\hat{F}_{i-1,p}$ be the accumulated importance of parameter p. The penalty is

$$\mathcal{L}_{\text{reg}}^{\text{EWC}} = \sum_p \hat{F}_{i-1,p} \left(\theta_{i,p} - \theta_{i-1,p}\right)^2.$$

3.3. Finalization Phase

After completing experience i, PoseAdapt computes and stores strategy-specific state: LFL and LwF update the teacher snapshot $\tilde{\mathcal{M}}_i$, whereas EWC computes fresh Fisher importances $F_{i,p}$ using a dedicated pass over $\mathcal{D}_i$:

$$F_{i,p} = \frac{1}{|\mathcal{B}_i|} \sum_{b \in \mathcal{B}_i} \left(\frac{\partial \mathcal{L}_{\text{kpt}}(b)}{\partial \theta_p} \right)^2,$$

then update the accumulated values via $\hat{F}_{i,p} = \lambda\, \hat{F}_{i-1,p} + F_{i,p}$. For the initial model $\mathcal{M}_0$, importances are computed from the COCO validation split as we assume no access to pretraining data. These finalization steps produce the reference state required for experience $i+1$ and make strategy behavior explicit and reproducible.

4. Benchmark and Experiments

PoseAdapt defines a controlled evaluation setting for continual adaptation in 2D human pose estimation with two complementary tracks: *domain-incremental* and *class-incremental*. The goal is to isolate the behavior of continual learning (CL) strategies under deployment-motivated constraints while holding capacity and data access fixed.

4.1. Experimental Setup

Methods. We evaluate *naïve fine-tuning* (FT), *Elastic Weight Consolidation* (EWC) [26], *Less-Forgetful Learning* (LFL) [24], and *Learning without Forgetting* (LwF) [30].

Figure 5. **Reference domain.** Examples from the COCO validation set representing the well-lit RGB baseline used for the initial experience in all benchmarks.

Reference model and data. All methods adapt the same off-the-shelf top-down pose estimator. To isolate keypoint regression from person detection, *ground-truth detection boxes* are used. The model (RTMPose-t [21], ∼3M parameters) is pretrained on the COCO [31] and AIC datasets to predict *17 keypoints* with *70.06 AP* on the COCO validation set[2]. We denote this as the reference domain (Fig. 5) in the first experience which the model must not forget as it adapts to new experiences with various domain shifts.

Metrics. The mean average precision (AP) at the end of training all experiences is reported. In addition, following standard continual learning evaluation [35], we report retention accuracy (RA) and average forgetting (AF). Let $a_{i,j}$ denote the AP on experience j after training on experience i. For a sequence of T experiences:

$$\mathrm{RA} = \tfrac{1}{T} \sum_{t=1}^{T} a_{T,t}, \tag{1}$$

$$\mathrm{AF} = \tfrac{1}{T-1} \sum_{t=1}^{T-1} \left(a_{t,t} - a_{T,t} \right). \tag{2}$$

RA measures how well the final model performs across all experiences, while AF quantifies the average drop in performance on earlier experiences after completing. When interpreted together, RA reflects overall stability, and AF indicates how much this stability depends on preserving early experiences (i.e., backward transfer). High RA combined with low AF indicates effective continual learning.

Hyperparameters and constraints. Adaptation uses AdamW [36] with a base learning rate 0.004, no weight decay, random seed 21, and a small linear warm-up. All images are normalized with $\mu = [123.675, 116.28, 103.53], \sigma = [58.395, 57.12, 57.375]$ and augmented using random horizontal flips and half-body occlusions. Each experience is allocated at most 1k labeled images and 10 epochs. Past data are never stored or revisited. The only state retained across experiences is the most recent network; distillation-based methods additionally keep a teacher snapshot from the previous experience. For EWC, the diagonal Fisher and parameter means are computed at the end of each experience on its training stream and stored for the next step.

[2] https://download.openmmlab.com/mmpose/v1/projects/rtmposev1/rtmpose-tiny_simcc-aic-coco_pt-aic-coco_420e-256x192-cfc8f33d_20230126.pth

4.2. Domain-Incremental Track

The domain-incremental track evaluates the ability to adapt across progressively more challenging shifts while retaining performance on previously seen domains. All domain shifts are generated post-hoc from the reference data to ensure reproducibility under shared identities and annotations.

4.2.1. Scene Density

Shift generation. Scene density shifts are created in two stages: image selection followed by synthetic occlusion. For each difficulty level, we first filter COCO images by the number of annotated people: 3–6, 7–10, and ≥11 individuals respectively (with minimum keypoint counts of 10/1/1). This yields increasingly crowded scenes while preserving original COCO annotations. Next, we apply fixed-budget cutout occlusion to each image: square blocks of random color are stamped at random locations such that approximately 5%, 10%, or 20% of the image area is occluded, using 10, 25, or 50 blocks per image. Together, these choices produce three reproducible density experiences—*O5*, *O10*, *O20*—with increasing scene density as shown in Fig. 6.

Figure 6. **Occlusion data.** Increasing scene density combined with synthetic cutout occlusion produces the three density-shift experiences: O5 (light), O10 (medium), and O20 (heavy).

Results. Across O5, O10, and O20, all methods exhibit moderate forgetting, with stability decreasing as crowding and occlusion intensify. In single-step adaptation, LwF is most stable for light occlusion (AF = 13.16) and maintains the highest RA there (56.88). Under medium and heavy occlusion, LFL becomes the most reliable, giving the lowest forgetting (AF = 13.58/14.23) and highest RA (50.41/52.82). In the sequential setting, accuracy drops compared to the pretrained reference (RA = 59.06), but LFL is consistently the strongest regularizer with RA = 51.02, while LwF achieves the lowest forgetting (AF = 5.97) but slightly lower RA.

DENSITY	Light				Medium				Heavy				Sequential					
Method	N	O5	AF ↓	RA ↑	N	O10	AF ↓	RA ↑	N	O20	AF ↓	RA ↑	N	O5	O10	O20	AF ↓	RA ↑
FT	53.92	54.18	16.14	54.05	52.92	42.41	17.15	47.66	51.90	48.70	18.16	50.30	52.35	53.39	42.54	49.69	6.16	49.49
EWC	53.04	52.25	17.02	52.65	53.23	40.94	16.83	47.09	51.78	47.42	18.29	49.60	51.81	53.15	41.89	48.86	6.13	48.93
LFL	56.78	56.78	13.29	56.78	**56.49**	**44.34**	**13.58**	**50.41**	**55.83**	**49.81**	**14.23**	**52.82**	**54.27**	54.96	**44.29**	**50.57**	6.31	**51.02**
LwF	**56.91**	**56.86**	**13.16**	**56.88**	54.97	42.92	15.10	48.94	54.81	49.61	15.25	52.21	53.92	**55.45**	43.33	49.80	**5.97**	50.62

LIGHTING	Low				Very Low				Extremely Low				Sequential					
Method	WL	LL	AF ↓	RA ↑	WL	VLL	AF ↓	RA ↑	WL	ELL	AF ↓	RA ↑	WL	LL	VLL	ELL	AF ↓	RA ↑
FT	53.83	51.45	16.24	52.64	48.51	46.74	21.56	47.62	27.55	40.07	42.52	33.81	36.99	40.29	43.10	39.09	15.86	39.87
EWC	52.63	50.15	17.43	51.39	45.92	43.21	24.15	44.56	20.74	40.64	49.32	30.69	28.65	36.84	41.92	39.45	19.51	36.72
LFL	**59.16**	**56.47**	**10.91**	**57.81**	**54.60**	**51.04**	**15.47**	**52.82**	**39.93**	43.22	**30.14**	**41.57**	**37.42**	**43.90**	**46.54**	**40.73**	**15.58**	**42.15**
LwF	58.37	55.74	11.69	57.06	52.07	49.00	18.00	50.53	38.87	**43.53**	31.20	41.20	26.17	34.01	39.71	37.51	24.01	34.35

MODALITY	Grayscale				Depth Image				Sequential				
Method	RGB	Gray	AF ↓	RA ↑	RGB	Depth	AF ↓	RA ↑	RGB	Gray	Depth	AF ↓	RA ↑
FT	48.25	50.70	21.82	49.47	6.66	35.43	63.40	21.05	5.45	4.90	35.42	55.17	15.26
EWC	47.87	48.92	22.19	48.39	**8.68**	36.17	**61.39**	**22.42**	**13.92**	**12.60**	35.19	**47.52**	**20.57**
LFL	**48.80**	**53.46**	**21.26**	**51.13**	6.72	32.74	63.34	19.73	12.48	9.53	35.13	50.33	19.05
LwF	46.68	50.55	23.39	48.61	5.41	**37.49**	64.66	21.45	13.23	10.29	**37.36**	50.14	20.30

Table 1. **PoseAdapt Domain-Incremental Results.** Accuracy (AP) of continual learning strategies across three domain-incremental benchmarks: Scene Density (increasing crowding/occlusion), *Lighting* (progressively darker conditions), and *Modality* (RGB → grayscale, depth, silhouette). Blue entries denote the reference/normal domain, while Red entries denote the target/shifted domains requiring adaptation. *RA* indicates the mean AP across listed domains at the end of training. FT is the naive finetuning. Because the benchmark enforces a fixed lightweight backbone and a strict adaptation budget, fine-tuning (FT) can underperform a frozen pretrained model, even on the latest domain. This is expected and highlights the difficulty of sustainable adaptation under constrained resources. Regularization-based methods are compared under single-step and sequential (online) protocols.

4.2.2. Lighting

Shift generation. Low-light images (*LL*) are obtained by scoring every COCO image (with 5+ annotated keypoints) using a brightness metric that averages (i) the global grayscale brightness of the full image and (ii) the grayscale brightness inside all person bounding boxes. This joint score emphasizes scenes where both the environment and the subjects are dark. The 1,000 images with the lowest brightness scores form the LL split. Progressively darker variants—*VLL* and *ELL*—are then synthesized from LL by applying controlled photometric degradations, primarily brightness reduction and contrast modification with very mild noise/blur, yielding increasing darkness levels shown in Fig. 7.

Low (LL) **Very Low** (VLL) **Extremely Low** (ELL)

Figure 7. **Lighting data.** This consists of three levels of decreasing illumination—LL, VLL, and ELL—obtained through brightness-based selection and controlled photometric darkening.

Results. Lighting shifts are considerably harder than density. In single-step adaptation, LFL again provides the best stability at every darkness level, with AF = 10.91/15.47/30.14 for LL/VLL/ELL. FT adapts aggressively to the new domain but heavily sacrifices the well-lit reference (e.g., WL→27.55 under ELL). LwF tracks LFL under LL but degrades more sharply at VLL and ELL. Sequentially, forgetting compounds strongly across LL→VLL→ELL. LFL remains the most robust method (RA = 42.15), with FT next (39.87), while EWC and LwF trail (36.72 and 34.35). The pattern reflects an increasingly nonlinear stability–plasticity trade-off as illumination diminishes.

4.2.3. Modality

Shift generation. Modality shifts comprise two experiences: *Gray* and *Depth*. Grayscale images are produced by fully desaturating each RGB frame and then applying a sequence of mild photometric perturbations including some Gaussian noise, increased contrast, and a small brightness correction. Depth images are generated by running MiDaS (DPT-Large) [48] per frame, resizing predictions back to the original resolution, and min–max normalizing the relative-depth map to [0,1]. Both outputs, illustrated in Fig. 8, are tiled to three channels for backbone compatibility.

Results. Modality shifts are the most severe. For grayscale, stability is modest: FT and EWC forget RGB strongly (AF ≈ 22), whereas LFL is most stable (AF = 21.26) and

Grayscale (Gray) **Depth Maps** (Depth)

Figure 8. **Modality data.** The modality benchmark includes Gray (desaturated and perturbed grayscale) and Depth (MiDaS-derived relative-depth maps scaled to [0, 1]), both tiled to three channels.

gives the highest Gray-domain AP (53.46). Depth presents an extreme shift: LwF yields the best Depth AP (37.49), while EWC retains RGB best (AF = 61.39, still catastrophic). Sequentially, no method maintains usable RGB performance when moving through Gray→Depth: RA collapses to 15–20 across all methods (best: EWC at 20.57). This confirms the large appearance and geometric mismatch between RGB images and monocular depth maps.

4.3. Class-Incremental Track

PoseAdapt supports skeleton growth scenarios analogous to class-incremental continual learning. New keypoints are introduced over time whereas labels for some past keypoints might not be available anymore. The model is expected to remember all seen keypoints while incorporating new ones to predict an expanded skeleton at each experience.

For standard evaluations of pure class-incremental continual pose estimation, we propose the PoseAdapt-BodyParts benchmark. In this benchmark, the network must progressively learn body, feet, hands, face, and spine keypoints. Body annotations with 17 keypoints from the COCO dataset [31], feet, face, and hand annotations with 6, 68, and 21 per hand keypoints respectively from the COCO-Wholebody dataset [22], and 9 spine annotations from SpineTrack [25] are used. Across the five experiences, the number of keypoints changes from 17, 23, 91, and 133 to 142 in the last experience. Training images from COCO dataset, shared by all three datasets, rule out any domain shifts. This allows the benchmark to solely assess skeleton growth capabilities of different continual learning strategies. As in the domain-incremental track, an off-the-shelf body pose estimator is used for the first experience, and the other four experiences are fine-tuned with maximum 10 epochs per experience.

Unlike the domain-incremental scenario, the model architecture must adapt to accommodate increasing number of keypoints. In regularization-based CL methods, this adaptation is limited to the head layers only, which are expanded to output the required number of channels while preserving learned weights of the existing channels. This setting tests the model's ability to extend its output space over time without losing earlier keypoint accuracy—a crucial requirement in custom-skeleton scenarios. Evaluation of this proposed class-incremental benchmark is left as a future work to maintain focus on domain shifts in this work.

4.4. Discussion

Across all domain-incremental tracks, the constrained 1k/10-epoch budget makes naïve FT highly unstable: it adapts strongly to the current domain but rapidly erodes earlier ones, often falling below the frozen reference model. The experience–experience matrices for FT in Figs. 9a to 9c make this clear—off-diagonal entries fade fast, especially under Lighting and Modality, where forgetting compounds after each shift.

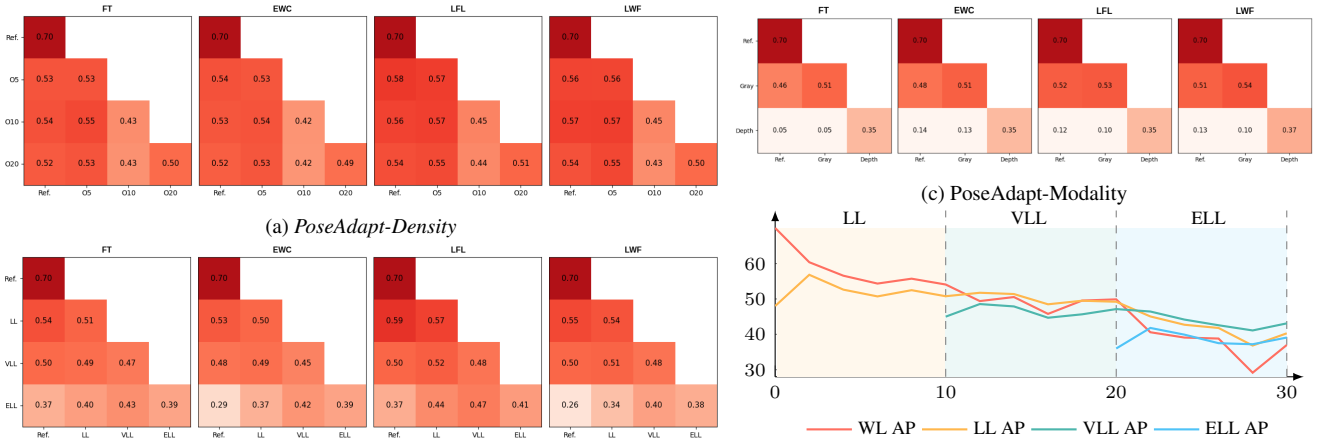
Among regularizers, **LFL** provides the most reliable stability under photometric changes. In both the tables and the heatmaps, its diagonals remain comparatively high across WL→LL→VLL→ELL, and its off-diagonals decay slowest. **LwF** handles mild shifts well and achieves the best single-step Depth AP, but its sequential matrices reveal larger cumulative drift. **EWC** retains earlier domains better in the Modality sequence, yet its diagonals under strong shifts are consistently lower, indicating limited plasticity.

The relative difficulty of the three benchmarks is also visible in the heatmaps: Density shows the mildest degradation, Lighting induces intermediate drift, and Modality exhibits the steepest collapse—particularly once Depth is introduced. The persistent RGB→Depth gap across all methods underscores that regularization alone is insufficient for cross-sensor adaptation. Overall, PoseAdapt exposes clear stability–plasticity trade-offs that become increasingly pronounced under severe distribution shifts.

5. Conclusion

We presented **PoseAdapt**, a standardized benchmark and toolkit for continual human pose estimation under controlled conditions: fixed lightweight backbones, no access to past data, and tightly bounded adaptation budgets. Evaluating FT, EWC, LFL, and LwF across density, lighting, and modality shifts, reveals consistent patterns: (1) FT is brittle and frequently underperforms the frozen reference; (2) regularization improves retention but is sensitive to shift severity; (3) LFL is the most stable across photometric domains, whereas LwF offers superior target-domain plasticity; and (4) none of the methods handle RGB→Depth robustly.

By unifying protocols, metrics (RA, AF), and shift generation pipelines, PoseAdapt establishes a reproducible foundation for studying sustainable adaptation in pose estimation.



(a) *PoseAdapt-Density*

(b) *PoseAdapt-Lighting*

(c) PoseAdapt-Modality

(d) Sequential adaptation curves for the lighting benchmark.

Figure 9. **Experience–experience performance matrices for the three domain-incremental benchmarks.** Heatmaps (a–c) visualize forgetting and retention across Density, Lighting, and Modality. Rows denote the training experience, columns the evaluation experience, and diagonal cells show immediate post-training AP. Off-diagonal decay illustrates how quickly methods forget earlier domains as adaptation progresses. Density produces only mild degradation, Lighting shows progressively stronger drift, and Modality exhibits the steepest collapse—especially after Depth. (d) plots the sequential WL→LL→VLL→ELL trajectory, highlighting the compounding loss on well-lit images as illumination decreases.

Impact. PoseAdapt provides a modular and reproducible framework that highlights concrete design targets for continual pose estimation: stronger feature alignment for modality changes, more stable regularizers for photometric shifts, and principled head-expansion strategies for skeleton growth. We hope the benchmark accelerates progress toward continual models suitable for long-term, real-world deployment.

Limitations. Most shifts are synthetic, and while they isolate controllable factors, they do not fully capture real-world sensor characteristics or motion-induced artefacts. The fixed-backbone assumption focuses the evaluation on adaptation strategies but excludes architectural innovations that may improve robustness to severe shifts. Finally, the benchmark considers only 2D single-frame keypoints and does not address temporal consistency or 3D pose estimation. Moreover, while skeleton growth scenarios are supported, they are not explicitly benchmarked here. Instead, this work focused on presenting the PoseAdapt framework and a systematic benchmark under high-stress conditions. Performance of continual pose estimation on a real-world application where domain and keypoint shifts both appear together is also not investigated.

Future Work. Future extensions should incorporate adapter-based or replay-informed continual learning, explore architectures with explicit cross-modal priors, and expand to transformers or bottom-up approaches. Extending PoseAdapt to video-based and 3D settings would allow systematic evaluation under temporal and geometric shifts. Broader domain coverage beyond synthetic shifts, such as clinical images or sports datasets, would further improve ecological validity.

Acknowledgement. The work leading to this publication was co-funded by the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101135724 (project LUMINOUS) and Grant Agreement No 101092889 (project SHARESPACE).

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1, 3

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018. 3

[3] Taravat Anvari and Kyoungju Park. 3d human body pose estimation in virtual reality: A survey. In *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 624–628. IEEE, 2022. 1

[4] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020. 3

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019. 2

[6] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9233–9242, 2020. 3

[7] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*, 2019. 3

[8] Rama Chellappa, Jiang Liu, Chun Pong Lau, and Prithviraj Dhar. Some challenges and solutions in data-driven ai. In *Computer Vision*, pages 1–17. Chapman and Hall/CRC, 2024. 1

[9] Junjie Chen, Weilong Chen, Yifan Zuo, and Yuming Fang. Recurrent feature mining and keypoint mixup padding for category-agnostic pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22035–22044, 2025. 2

[10] Gaetano Dibenedetto. A new perspective in health recommendations: Integration of human pose estimation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 1382–1387, 2024. 1

[11] Arthur Douillard and Timothée Lesort. Continuum: Simple management of complex continual learning scenarios. *arXiv preprint arXiv:2102.06253*, 2021. 3

[12] Michael Essich, Markus Rehmann, and Cristóbal Curio. Auxiliary task-guided cyclegan for black-box model domain adaptation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 541–550, 2023. 1

[13] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 3

[14] Zheyan Gao, Jinyan Chen, Yuxin Liu, Yucheng Jin, and Dingxiaofei Tian. A systematic survey on human pose estimation: upstream and downstream tasks, approaches, lightweight models, and prospects. *Artificial Intelligence Review*, 58(3):68, 2025. 1

[15] Yanan Gu, Xu Yang, Kun Wei, and Cheng Deng. Not just selection, but exploration: Online class-incremental continual learning via dual view consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7442–7451, 2022. 3

[16] Romain Guesdon, Carlos Crispim-Junior, and Laure Tougne. Dripe: A dataset for human pose estimation in real-world driving settings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2865–2874, 2021. 1

[17] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. 3

[18] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024. 1

[19] Rongtian Huo, Qing Gao, Jing Qi, and Zhaojie Ju. 3d human pose estimation in video for human-computer/robot interaction. In *International conference on intelligent robotics and applications*, pages 176–187. Springer, 2023. 1

[20] Christian Keilstrup Ingwersen, Christian Møller Mikkelstrup, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjorholm Dahl. Sportspose-a dynamic 3d sports pose dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5219–5228, 2023. 1

[21] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 1, 2, 5

[22] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 7

[23] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 1

[24] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. *arXiv preprint arXiv:1607.00122*, 2016. 3, 4

[25] Muhammad Saif Ullah Khan, Stephan Krauß, and Didier Stricker. Towards unconstrained 2d pose estimation of the human spine. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6172–6181, 2025. 1, 3, 7

[26] 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. 2, 3, 4

[27] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021. 3

[28] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. *Advances in neural information processing systems*, 30, 2017. 3

[29] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for hu-

man pose estimation. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 1

[30] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2, 3, 4

[31] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3, 5, 7

[32] Yinglu Liu, Hailin Shi, Hao Shen, Yue Si, Xiaobo Wang, and Tao Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In *AAAI*, pages 11637–11644, 2020. 1

[33] 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

[34] Vincenzo Lomonaco, Lorenzo Pellegrini, Andrea Cossu, Antonio Carta, Gabriele Graffieti, Tyler L. Hayes, Matthias De Lange, Marc Masana, Jary Pomponi, Gido van de Ven, Martin Mundt, Qi She, Keiland Cooper, Jeremy Forest, Eden Belouadah, Simone Calderara, German I. Parisi, Fabio Cuzzolin, Andreas Tolias, Simone Scardapane, Luca Antiga, Subutai Amhad, Adrian Popescu, Christopher Kanan, Joost van de Weijer, Tinne Tuytelaars, Davide Bacciu, and Davide Maltoni. Avalanche: an end-to-end library for continual learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[35] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 5

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[37] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2637–2646, 2022. 1

[38] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018. 3

[39] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European conference on computer vision (ECCV)*, pages 67–82, 2018. 3

[40] Alexander Marusov, Mariam Kaprielova, and Radoslav Neychev. Enhancing human pose estimation with privileged learning. In *2022 31st Conference of Open Innovations Association (FRUCT)*, pages 174–180. IEEE, 2022. 1

[41] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989. 3

[42] Umberto Michieli and Pietro Zanuttigh. Incremental learning techniques for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 3

[43] MMPose Contributors. OpenMMLab Pose Estimation Toolbox and Benchmark, 2020. 2

[44] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[45] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019. 3

[46] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 3

[47] Miroslav Purkrabek and Jiri Matas. Probpose: A probabilistic approach to 2d human pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27124–27133, 2025. 1

[48] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 6

[49] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 3

[50] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 3

[51] István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3d human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2956–2966, 2023. 3

[52] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pages 4548–4557. PMLR, 2018. 3

[53] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 3

[54] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE international conference on computer vision*, pages 3400–3409, 2017. 3

[55] Hong Son Nguyen, DaEun Cheong, Andrew Chalmers, Myoung Gon Kim, Taehyun Rhee, and JungHyun Han. Interaction with virtual objects using human pose and shape estimation. *Computer Animation and Virtual Worlds*, 36(3): e70046, 2025. 1

[56] Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022. 3

[57] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11060–11068, 2022. 1

[58] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 3

[59] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt. Estimating egocentric 3d human pose in the wild with external weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13157–13166, 2022. 1

[60] Jingbo Wang, Ye Yuan, Zhengyi Luo, Kevin Xie, Dahua Lin, Umar Iqbal, Sanja Fidler, and Sameh Khamis. Learning human dynamics in autonomous driving scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20796–20806, 2023. 1

[61] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 2

[62] 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. 1, 2

[63] Calvin Yeung, Tomohiro Suzuki, Ryota Tanaka, Zhuoer Yin, and Keisuke Fujii. Athletepose3d: A benchmark dataset for 3d human pose estimation and kinematic validation in athletic movements. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5945–5956, 2025. 1

[64] Xiangyu Yin, Boyuan Yang, Weichen Liu, Qiyao Xue, Abrar Alamri, Goeran Fiedler, and Wei Gao. Progait: A multi-purpose video dataset and benchmark for transfemoral prosthesis users. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8984–8993, 2025. 1

[65] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017. 3

[66] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102, 2020. 1

[67] Jingxiao Zheng, Xinwei Shi, Alexander Gorban, Junhua Mao, Yang Song, Charles R Qi, Ting Liu, Visesh Chari, Andre Cornman, Yin Zhou, et al. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4478–4487, 2022. 1

[68] Zirui Zhou, Junhao Liang, Zizhao Peng, Chao Fan, Fengwei An, and Shiqi Yu. Gait patterns as biomarkers: A video-based approach for classifying scoliosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 284–294. Springer, 2024. 1