DualAlign: A Heterogeneous Distillation Framework for Extremely Lightweight Pose Estimation

Liyang Zheng*, Miao Zhang*, Siyuan Fang†, Xing Chen†
*Harbin Institute of Technology, Shenzhen, China
†HeyiSpace, Shenzhen, China
Email: liyangzheng8@gmail.com, zhangmiao@hit.edu.cn
fangsiyuan@szh1.com.cn, chenxing@szh1.com.cn

***Abstract*—Deploying Human Pose Estimation (HPE) on microcontroller units (MCUs) often relies on hardware-aware architecture selection (e.g., XiNet-Pose) to meet strict resource constraints, which inevitably sacrifices performance by excluding heavy components like Spatial Pyramid Pooling (SPP) and multi-scale heads. While knowledge distillation (KD) offers a remedy, it falters in such highly heterogeneous scenarios due to *Optimization Misalignment* from conflicting gradient directions and *Structural Misalignment* between multi-head teachers and single-head students. To bridge these gaps, we propose DualAlign, a unified framework for extremely lightweight pose estimation. DualAlign incorporates two complementary alignment mechanisms: Gradient-Projected Alignment (GPA) dynamically gates distillation weights based on gradient directional consistency to automatically shield the student from optimization conflicts, while Transient Structural Alignment (TSA) employs auxiliary heads governed by a "Lead Head Principle" to seamlessly inject multi-scale knowledge into the shared backbone during training. Extensive experiments on MS COCO Keypoints demonstrate that DualAlign significantly recovers accuracy without incurring any inference overhead, consistently outperforming other representative distillation paradigms adapted for this constrained setting.**

***Index Terms*—TinyML, Knowledge Distillation, Human Pose Estimation**

I. Introduction

Real-time Human Pose Estimation (HPE) on edge devices, particularly microcontroller units (MCUs), is pivotal for ubiquitous applications ranging from health monitoring to human-machine interaction [1]. However, deploying deep learning models on such strictly constrained hardware (e.g., typically $<$ 1MB RAM) necessitates hardware-aware architecture design [2]. To meet these stringent resource budgets, recent lightweight models often undergo aggressive structural exclusion, where computationally heavy but semantic-rich components—such as Spatial Pyramid Pooling (SPP) and multi-scale detection heads—are removed during the architecture selection phase [2]. While this simplification ensures inference efficiency, it inevitably compromises the model's representational capacity, leading to a significant degradation in pose estimation accuracy, especially under occlusion or scale variation.

Corresponding author.

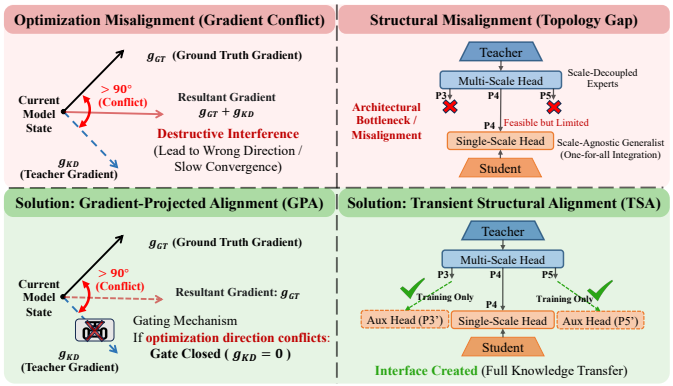


Fig. 1. Illustration of the challenges and our proposed solutions in heterogeneous distillation.(Top) Optimization Misalignment occurs when regression gradient directions conflict, causing destructive interference; Structural Misalignment arises from the topology gap, where the single-head student lacks interfaces to receive multi-scale knowledge. (Bottom) Our DualAlign framework resolves these via GPA (dynamic gating based on gradient consistency) and TSA (transient auxiliary heads), respectively.

Knowledge Distillation (KD) [3] serves as a promising avenue to recover this lost accuracy by transferring knowledge from a high-capacity teacher to the compact student. In this context, employing the full-architecture supernet as the teacher is the most natural choice, as it maximizes the alignment of inductive biases (e.g., convolutional operators) with the student [4]. However, even within this same-lineage pair, the structural exclusion creates a scenario of *highly heterogeneous distillation*. We observe that conventional KD methods [5], [6], which typically rely on static weights or heuristic loss-based gating, falter in this extreme regime. They fail to address two fundamental misalignments arising from the structural gap between the full-capacity teacher and the structurally simplified student.

The first obstacle is **Optimization Misalignment**, inherent in distilling regression-based tasks with unbounded outputs. As illustrated in Fig. 1 (Top-Left), lacking global context components (e.g., SPP), the student's optimization direction often

contradicts the teacher's. When the teacher's gradient (g_{KD}) diverges significantly from the ground truth gradient (g_{GT}) (i.e., exhibiting *negative cosine similarity*), blindly mimicking such conflicting signals destabilizes training. This creates a critical need to dynamically identify *when the teacher's guidance constructively aligns with the ground truth* versus when it constitutes harmful noise.

The second obstacle is **Structural Misalignment**, stemming from the topological discrepancy between the models. As depicted in Fig. 1 (Top-Right), while the teacher utilizes a multi-level feature pyramid (P3-P5) to handle scale variations, the hardware-constrained student is typically restricted to a Single-Head architecture (e.g., P4 only). This creates an *architectural bottleneck*: the student lacks the corresponding physical interfaces to receive the teacher's distributed scale-specific knowledge (e.g., P3 for small objects), leaving these knowledge channels disconnected (marked as "No Interface"). This necessitates a strategy to determine *how to effectively inject multi-scale knowledge into a simplified single-scale structure* despite the architectural gap.

To bridge these gaps, we propose **DualAlign**, a unified distillation framework tailored for extremely lightweight pose estimation models (e.g., XiNet-pose [2]). DualAlign incorporates two complementary mechanisms to resolve the aforementioned misalignments without introducing any inference overhead (Fig. 1, Bottom). First, to ensure optimization consistency, we introduce **Gradient-Projected Alignment (GPA)**. Instead of relying on scalar loss values, GPA monitors the gradient directional consistency between the teacher and the student. It dynamically gates the distillation signal, acting as an automatic "safe valve" that allows knowledge transfer only when the teacher's guidance constructively aligns with the ground-truth optimization direction. Second, to ensure structural consistency, we propose **Transient Structural Alignment (TSA)**. During training, we construct auxiliary heads for the student, governed by a "Lead Head Principle", to establish a temporary isomorphic structure with the teacher. This allows the seamless injection of multi-scale knowledge into the shared backbone, which is then retained by the single lead head during inference.

Our main contributions are summarized as follows:

- We systematically identify Optimization Misalignment and Structural Misalignment as the primary causes of failure for heterogeneous distillation in extremely lightweight regression tasks.
- We propose Gradient-Projected Alignment (GPA), a hyperparameter-free gating mechanism that resolves gradient conflicts with automatic curriculum learning.
- We design Transient Structural Alignment (TSA) to bridge the topological gap between multi-head teachers and single-head students, enabling effective feature injection without inference costs.
- Extensive experiments on MS COCO Keypoints show that DualAlign significantly recovers the performance of structurally simplified models, outperforming representative distillation methods in heterogeneous scenarios.

II. Related Work

A. Efficient HPE and Knowledge Distillation

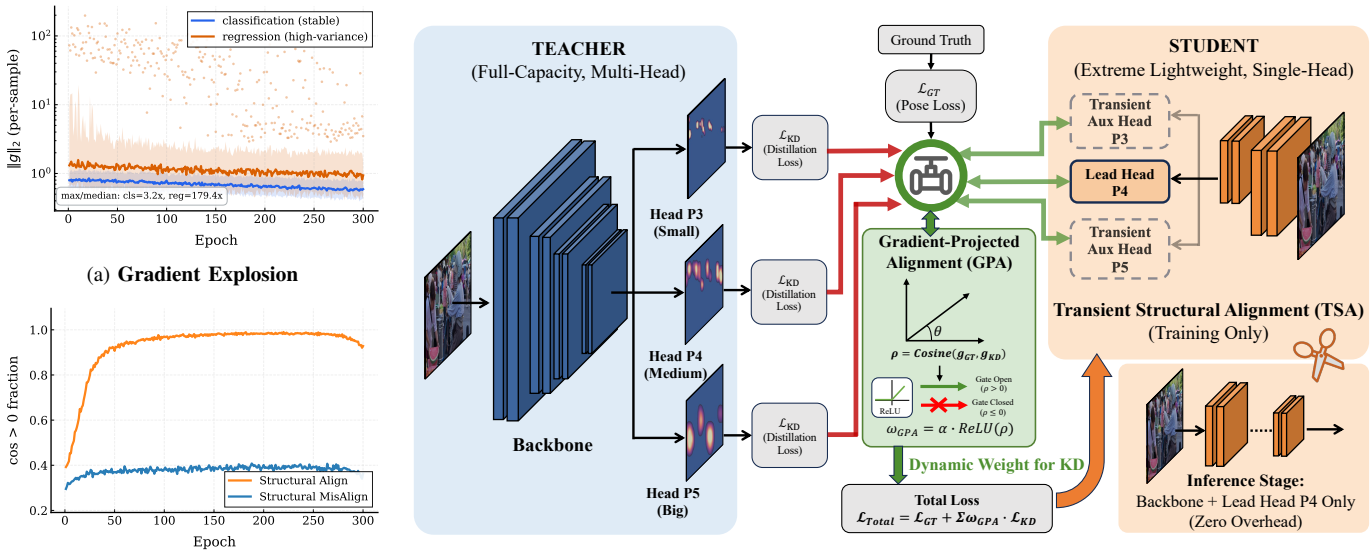
To enable real-time HPE on edge devices, recent works have shifted from computation-heavy heatmap-based paradigms [7], [8] to regression-based architectures [9]–[11]. While removing heatmaps and post-processing significantly reduces latency, deploying on MCUs requires further hardware-aware architectural selection (e.g., XiNet-Pose [12]), often resulting in the exclusion of critical components like SPP and multi-scale heads. KD [3], [13], [14] is a standard approach to recover such capacity loss. However, conventional feature-based methods [15]–[17] typically require computationally expensive alignment layers (e.g., 1×1 convs) to bridge the dimension gap. Moreover, directly applying response-based KD to regression tasks often yields sub-optimal results due to the unbounded nature of coordinate prediction and the lack of spatial regularization inherent in heatmaps [18]–[20], necessitating a more robust distillation strategy for heterogeneous regression pairs.

B. Addressing Heterogeneity and Conflicts

The distillation process between a full-capacity teacher and a structurally simplified student inherently constitutes a severely heterogeneous scenario. Such architectural divergence fundamentally creates two primary obstacles: optimization conflicts and structural misalignment. To mitigate these issues, prior works typically resort to adaptive reweighting strategies [5], [6], [21]–[23] to balance conflicting supervision signals, or introduce architectural adaptation modules [15], [16], [24]–[26] to bridge the feature dimension gap.

Optimization Conflict Resolution. Existing methods often employ heuristic gating to mitigate optimization conflicts. GID [5] and FGD [6] select distillation samples based on *scalar loss discrepancies* (e.g., distilling only when teacher loss is lower). We argue that scalar loss is a lossy proxy for optimization dynamics, as it ignores the *direction* of parameter updates. While gradient-guided methods like MoKD [23] attempt to project conflicting gradients, they often overlook the *gradient magnitude explosion* [20] inherent in regression tasks, which can destabilize training. In contrast, our Gradient-Projected Alignment (GPA) utilizes normalized cosine similarity to monitor gradient directional consistency, acting as a robust gate that filters out conflicts regardless of magnitude fluctuations.

Structural Alignment. To bridge the topological gap, methods like HEAD [25] introduce assistant heads to align backbone features. However, HEAD targets general object detection and does not address the specific *Task Assignment Conflict* in TinyML, where a multi-head teacher distributes objects across scales while a single-head student must handle all scales. Our Transient Structural Alignment (TSA) extends this concept by enforcing a "Lead Head Principle", which creates a temporary isomorphic structure to inject multi-scale knowledge into the shared backbone without affecting the inference-time single-head topology.



(a) **Gradient Explosion**

(b) **Structural Bottleneck**

(c) **The DualAlign Framework.**

Fig. 2. Motivation and Overview of DualAlign. (a) Optimization Misalignment: Regression gradients (orange) exhibit extreme magnitude variance (log-scale) compared to stable classification gradients (blue), risking training collapse. (b) Structural Misalignment: In heterogeneous settings (e.g., Multi-Head → Single-Head), the fraction of constructive gradients ($\cos > 0$) remains persistently low (Blue line), indicating a closed learning gate. Structural alignment significantly improves gradient consistency (Orange line). (c) Methodology: To resolve these, GPA dynamically gates gradients based on directional consistency ρ, while TSA introduces transient auxiliary heads to bridge the topological gap. The auxiliary heads are discarded during inference.

III. Methodology

In this section, we formulate the distillation problem specifically for heterogeneous HPE. We first present our empirical analysis (Fig. 2a and Fig. 2b), which reveals the dual challenges of optimization conflict and structural bottlenecks. Based on these insights, we introduce DualAlign, a unified framework comprising GPA and TSA. The overall framework is illustrated in Fig. 2c.

A. Preliminaries

Let $\mathcal{T}$ and $\mathcal{S}$ denote the teacher and student networks, parameterized by $\theta_{\mathcal{T}}$ and $\theta_{\mathcal{S}}$, respectively. We adopt the XiNet-pose family [2], where $\mathcal{T}$ is the full-capacity supernet equipped with multi-scale heads (P3, P4, P5), and $\mathcal{S}$ is a structurally simplified subnet constrained to a single head (Lead P4) to minimize peak memory on MCUs.

Given an input image X and its corresponding ground truth labels Y (containing bounding boxes B and keypoints K), the student is trained using a task loss $\mathcal{L}_{GT}$ comprising classification (BCE), bounding box regression (IoU/DFL), and keypoint regression (OKS/MSE). KD introduces an additional loss $\mathcal{L}_{KD}$ to enforce consistency between the predictions of $\mathcal{S}(X)$ and $\mathcal{T}(X)$. The generic optimization objective to find the optimal student parameters $\theta^*_{\mathcal{S}}$ is formulated as:

$$\theta^*_{\mathcal{S}} = \arg\min_{\theta_{\mathcal{S}}} \left(\mathcal{L}_{GT}(\mathcal{S}(X;\theta_{\mathcal{S}}), Y) + \lambda \mathcal{L}_{KD}(\mathcal{S}, \mathcal{T}) \right), \quad (1)$$

where λ is typically a static balancing weight.

B. Motivation and Empirical Analysis

Standard KD fails in highly heterogeneous regression tasks. We investigate the underlying causes through gradient analysis, identifying two critical misalignments.

a) Optimization Misalignment: The Gradient Conflict.: Regression tasks inherently suffer from high variance in gradient magnitudes. As visualized in Fig. 2a, regression gradients (orange curve) exhibit extreme magnitude explosions, often spanning orders of magnitude higher than stable classification gradients (blue curve). When the teacher's prediction contains bias, this high-variance regression gradient g_{KD} often contradicts the ground-truth gradient g_{GT}. Blindly aggregating these conflicting signals with a static weight leads to destructive interference. This necessitates a gating mechanism (GPA) to filter out harmful gradients based on *directional consistency* rather than unstable magnitudes.

b) Structural Misalignment: The Topological Bottleneck.: While gating prevents collapse, it exposes a deeper structural issue. We monitor the fraction of "constructive gradients" (where cosine similarity $\rho > 0$) during training. As shown in Fig. 2b, in the heterogeneous setting (Blue line), the positive fraction remains low (≈ 0.4), implying that the single-head student fundamentally struggles to assimilate the multi-head teacher's knowledge, causing the gate to remain "closed" most of the time. In contrast, structurally aligned settings (Orange line) achieve high consistency. This motivates us to construct a structural bridge (TSA) that re-aligns the feature topology to unlock valid knowledge transfer.

C. Gradient-Projected Alignment (GPA)

To address Optimization Misalignment, we propose GPA, a dynamic weighting strategy that governs *when to learn*. Instead of relying on scalar loss discrepancies, GPA monitors the geometric alignment of optimization directions. Let g_{GT} and g_{KD} denote the gradients of the task loss and distillation

loss w.r.t. the student's head parameters. We compute the normalized cosine similarity ρ:

$$\rho = \frac{g_{GT} \cdot g_{KD}}{\|g_{GT}\|\|g_{KD}\| + \epsilon}. \quad (2)$$

Normalization is crucial here to decouple the update direction from the magnitude explosions identified in Sec. III-B. The dynamic distillation weight ω_{GPA} is then formulated as:

$$\omega_{GPA} = \alpha \cdot \text{ReLU}(\rho), \quad (3)$$

where α is a base scaling factor. The ReLU operation acts as a hard filter: it zeroes out the distillation weight when $\rho < 0$ (conflict), preventing negative transfer. Conversely, when $\rho > 0$ (alignment), the weight scales with consistency. This mechanism naturally achieves automatic curriculum learning: early random gradients are suppressed, while converged gradients are amplified.

D. Transient Structural Alignment (TSA)

To address Structural Misalignment and "open the gate" for GPA, we propose TSA, answering *how to learn*. The core intuition is to construct a temporary isomorphic architecture that aligns the student's topology with the multi-head teacher during training. Given a teacher with heads $\{H_T^{P3}, H_T^{P4}, H_T^{P5}\}$ and a student restricted to a single head H_S^{P4}, we augment the student with Transient Auxiliary Heads $\{H_{aux}^{P3}, H_{aux}^{P5}\}$ attached to the shared backbone. To prevent the auxiliary heads from diluting the main task capability of the student, we enforce the **Lead Head Principle**, which decouples task execution from knowledge injection:

1) Lead-Exclusive Supervision (Task Execution): The ground-truth loss $\mathcal{L}_{GT}$ is applied *exclusively* to the student's Lead Head (H_S^{P4}). Unlike conventional multi-head training where objects are assigned to specific levels based on scale, here the single Lead Head is forced to handle objects of *all scales*. This ensures that H_S^{P4} develops the standalone capability required for inference.
2) Auxiliary-Only Distillation (Knowledge Injection): The Auxiliary Heads $\{H_{aux}^{P3}, H_{aux}^{P5}\}$ are supervised *solely* by the teacher via distillation loss $\mathcal{L}_{KD}$, without seeing any ground truth. They force the shared backbone to encode multi-scale semantics (e.g., small object details from P3) that match the teacher's representation.

Through this mechanism, the Lead Head "harvests" the enriched multi-scale features injected by the Auxiliary Heads, achieving performance gains without structural changes at inference time.

E. Unified Training Objective

Combining GPA and TSA, the final objective function is:

$$\mathcal{L}_{total} = \mathcal{L}_{GT}(H_S^{lead}, Y) + \sum_{k \in \{3,4,5\}} \omega_{GPA}^{(k)} \cdot \mathcal{L}_{KD}(H_S^{(k)}, H_T^{(k)}), \quad (4)$$

where $\omega_{GPA}^{(k)}$ indicates that gradient gating is computed independently for each head pair. Upon training completion, auxiliary heads are discarded.

IV. Experiments

A. Experimental Setup

Datasets and Metrics. We evaluate our method on the MS COCO Keypoints dataset, using the standard train2017 split for training and val2017 for evaluation. We report Average Precision (AP), AP_{50}, AP_{75}, and Average Recall (AR) based on Object Keypoint Similarity (OKS). All AP/AR metrics are reported in percentage (%).

Model Configurations. We target the heterogeneity challenges in compressing YOLO-like anchor-free pose estimators.

- **Student ($\mathcal{S}$):** We adopt a hardware-constrained version of XiNet-Pose [12], a YOLO-based architecture optimized for extreme edge devices. To minimize peak memory usage and latency, the student is structurally simplified to a single-head (P4) configuration with the Spatial Pyramid Pooling (SPP) neck module removed.
- **Teacher ($\mathcal{T}$):** We select the full-capacity version of XiNet-Pose as the teacher, equipped with SPP and multi-scale heads (P3, P4, P5). While other SOTA models (e.g., HRNet [7], ViTPose [27]) exist, we deliberately choose this homologous teacher to ensure the alignment of inductive biases (e.g., convolutional operators and regression logic). This minimizes the feature domain gap, ensuring that optimization conflicts arise solely from structural capacity constraints rather than fundamental representation mismatches [4].

Implementation Details. Following the standard protocol of YOLO-Pose [9], all models are trained for 300 epochs with a cosine learning rate schedule. We use the SGD optimizer and standard data augmentation strategies (e.g., Mosaic). For our proposed DualAlign, the base distillation weight α is set to 5.0. The Transient Auxiliary Heads are instantiated to mirror the teacher's corresponding heads (P3/P5) structurally but are strictly discarded during inference.

TABLE I
Main Results on COCO Val. Comparison with representative distillation paradigms adapted to the TinyML setting.

Method	Mechanism	AP (%)	AP_{50} (%)	AP_{75} (%)	AR (%)
Teacher (Full Arch)	-	48.3	77.4	51.2	72.2
Student (Baseline)	-	27.7	61.6	20.7	59.9
Static KD	Fixed Weight	26.1	60.3	18.4	60.0
GID [5]	Loss Gating	27.6	61.3	20.3	59.9
MoKD [23]	Gradient Proj.	25.5	59.4	19.4	58.2
HEAD [25]	Structural Aux	17.6	36.5	8.7	48.4
DualAlign (Ours)	**GPA + TSA**	**28.9**	**63.3**	**22.6**	**60.1**

B. Main Results

We evaluate DualAlign against representative distillation baselines adapted to the TinyML setting: Static KD (fixed weights), GID [5] (loss-based gating), MoKD [23] (gradient projection), and HEAD [25] (structural assistants).

As shown in Tab. I, directly applying static regression distillation causes negative transfer (-1.6 AP), suggesting that unfiltered teacher signals can conflict with task optimization in this heterogeneous regime. GID yields only marginal change (-0.1 AP), indicating that scalar-loss gating is insufficient

to reliably detect harmful updates. MoKD further degrades performance (-2.2 AP), consistent with instability under large-magnitude regression gradients. HEAD collapses severely (-10.1 AP), highlighting that structural adaptation without conflict-aware optimization can distort the student's feature learning.

In contrast, DualAlign is the only method with consistent positive gain (+1.2 AP over the baseline), and it also improves AP_{50} and AP_{75} simultaneously. This result supports our core claim: in extremely lightweight heterogeneous distillation, optimization conflict and structural mismatch must be addressed jointly rather than independently. Notably, DualAlign narrows the gap to the full-architecture teacher from 20.6 AP (Teacher–Student baseline) to 19.4 AP, showing that meaningful recovery is achievable even under strict TinyML constraints and without any extra inference-time components.

TABLE II
ABLATION STUDY ON COCO VAL. WE ANALYZE THE IMPACT OF CHANGING THE STUDENT STRUCTURE AND THE GATING STRATEGY.

Exp	Structure	Gating Strategy	AP (%)	AP_{50} (%)	AP_{75} (%)
1	Single-Head (Base)	*None*	27.7	61.6	20.7
2	Single-Head	Static	26.1	60.3	18.4
3	Single-Head	**GPA (Ours)**	28.0	62.1	21.0
4	TSA (Aux Heads)	Static	28.1	62.1	20.8
5	**TSA (Aux Heads)**	**GPA (Ours)**	**28.9**	**63.3**	**22.6**

C. Ablation Study

To validate each component, we vary **Structure** (Single-Head vs. TSA) and **Gating Strategy** (Static vs. GPA), as summarized in Tab. II.

Impact of GPA. Within the Single-Head setting, static regression KD degrades performance (Row 2 vs. Row 1), whereas replacing static weighting with GPA restores and slightly improves AP (Row 3). This confirms that directional-consistency gating effectively filters conflicting gradients and prevents destructive transfer.

Impact of TSA and Synergy. Adding TSA with static weighting (Row 4) provides only limited improvement, indicating that structural interfaces alone do not fully solve optimization conflict. When TSA is combined with GPA (Row 5), performance increases to the best overall result (+1.2 AP vs. baseline), demonstrating that TSA and GPA are complementary: TSA opens knowledge pathways, while GPA stabilizes their optimization dynamics. This pattern also suggests a practical training recipe: first ensure optimization safety through conflict-aware gating, then leverage structural augmentation to maximize transferable multi-scale knowledge.

D. In-Depth Analysis

To move beyond performance metrics and understand the internal working mechanism of DualAlign, we analyze the training dynamics from optimization and structural perspectives. While the gains in Sec. I and Tab. II verify effectiveness, they do not directly reveal *why* DualAlign remains stable under severe heterogeneity. We therefore examine internal optimization signals and structural consistency statistics to provide mechanistic evidence for the two alignment modules.

Dynamics of Gradient-Projected Alignment. One might question whether the benefit of GPA comes solely from implicitly finding an optimal static scalar. In Fig. 3, we visualize the evolution of the raw distillation weight ω_{raw} (averaged per epoch). We observe two critical behaviors that static methods cannot replicate:

- **Micro-level Filtering:** Although the average weight appears low ($\approx$ 0.3), this is a result of selective gating. As analyzed in Sec. III-B, regression gradients are highly unstable. GPA assigns high weights to constructive batches while strictly zeroing out conflicting ones. A static weight of 0.3 would enforce distillation on all batches, accumulating noise, whereas GPA acts as a noise filter.
- **Macro-level Auto-Annealing:** The weight curve exhibits a distinct decay trend in the late training phase (200+ epochs). As the student converges, GPA automatically reduces the supervision intensity, preventing the frozen teacher from over-regularizing the student's fine-grained optimization.

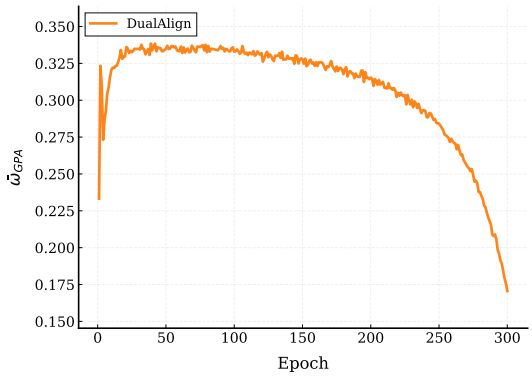


Fig. 3. Evolution of GPA Weight. The dynamic weight exhibits an "auto-annealing" trajectory, naturally decaying as the student converges, reducing conflict without manual scheduling.

Verification of Structural Conflict. The necessity of our gating mechanism is further supported by the gradient consistency statistics. As illustrated in the Motivation (Fig. 2b), in highly heterogeneous pairs (e.g., Multi-Head $\rightarrow$ Single-Head), the Cosine Positive Fraction—the ratio of batches where teacher and student gradients align—remains persistently low ($\approx$ 40%). This indicates that due to the topological gap, the teacher's guidance contradicts the student's ground-truth optimization direction in the majority of cases. By detecting this low positive fraction, GPA effectively shuts the gate for the conflicting 60% of iterations, protecting the student from destructive interference that causes the collapse observed in the Static KD baseline (Tab. II).

V. CONCLUSION

In this paper, we addressed the challenge of knowledge distillation for extremely lightweight pose estimation, where hardware-aware architecture selection creates a severe heterogeneity gap between full-capacity teachers and structurally simplified students. We systematically identified two

fundamental barriers: *Optimization Misalignment* caused by conflicting regression gradients, and *Structural Misalignment* caused by topological mismatches. To bridge these gaps, we proposed DualAlign, a unified framework that synergizes Gradient-Projected Alignment (GPA) for dynamic conflict suppression and Transient Structural Alignment (TSA) for effective feature injection. Extensive experiments on MS COCO Keypoints demonstrate that DualAlign significantly recovers the performance lost to architectural exclusion, surpassing existing paradigms for heterogeneous distillation in constrained scenarios. Crucially, our method achieves these gains with zero inference overhead, making it an ideal solution for deployment on microcontrollers.

ACKNOWLEDGMENT

The authors acknowledge the use of artificial intelligence (AI) to assist with English language editing and polishing. The authors certify that the content of this paper is original and reflects their own work.

REFERENCES

[1] N. Hoefflin, T. Spulak, A. Jeworutzki, and J. Schwarzer, "Real-time lateral sitting posture detection using yolov5," in *2024 10th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, pp. 711–715, IEEE, 2024.

[2] A. Ancilotto, F. Paissan, and E. Farella, "Xinet-pose: Extremely lightweight pose detection for microcontrollers," in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1–6, IEEE, 2024.

[3] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[4] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1365–1374, 2019.

[5] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, and E. Zhou, "General instance distillation for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7842–7851, 2021.

[6] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, and C. Yuan, "Focal and global knowledge distillation for detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4643–4652, 2022.

[7] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

[8] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 466–481, 2018.

[9] D. Maji, S. Nagori, M. Mathew, and D. 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*, pp. 2637–2646, 2022.

[10] Y. Li, S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, W. Yang, and S.-T. Xia, "Simcc: A simple coordinate classification perspective for human pose estimation," in *European Conference on Computer Vision*, pp. 89–106, Springer, 2022.

[11] W. McNally, K. Vats, A. Wong, and J. McPhee, "Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation," in *European Conference on Computer Vision*, pp. 37–54, Springer, 2022.

[12] A. Ancilotto, F. Paissan, and E. Farella, "Xinet: Efficient neural networks for tinyml," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16968–16977, 2023.

[13] A. M. Mansourian, R. Ahmadi, M. Ghafouri, A. M. Babaei, E. B. Golezani, Z. Y. Ghamchi, V. Ramezanian, A. Taherian, K. Dinashi, A. Miri, *et al.*, "A comprehensive survey on knowledge distillation," *arXiv preprint arXiv:2503.12067*, 2025.

[14] Z. Ji, W. Zhang, S. Qiao, K. Feng, and Y. Qian, "A coarse-to-fine human pose estimation method based on two-stage distillation and progressive graph neural network," *arXiv preprint arXiv:2508.11212*, 2025.

[15] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *In Proceedings of ICLR*, 2015.

[16] J. Guo, K. Han, Y. Wang, H. Wu, X. Chen, C. Xu, and C. Xu, "Distilling object detectors via decoupled features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2154–2164, 2021.

[17] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5311–5320, 2021.

[18] S. Ye, Y. Zhang, J. Hu, L. Cao, S. Zhang, L. Shen, J. Wang, S. Ding, and R. Ji, "Distilpose: Tokenized pose regression with heatmap distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2163–2172, 2023.

[19] Z. Zhengetal, "Localizationdistillationforobjectdetection," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 45, no. 8, pp. 10070–10083, 2023.

[20] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, and C. Lu, "Human pose regression with residual log-likelihood estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11025–11034, 2021.

[21] M. Zhang, L. Wang, D. Campos, W. Huang, C. Guo, and B. Yang, "Weighted mutual learning with diversity-driven model compression," *Advances in Neural Information Processing Systems*, vol. 35, pp. 11520–11533, 2022.

[22] Q. Lan and Q. Tian, "Gradient-guided knowledge distillation for object detectors," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 424–433, 2024.

[23] Q. Lan and Q. Tian, "Gradient-guided knowledge distillation for object detectors," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 424–433, 2024.

[24] L. Wang, C.-Y. Lee, Z. Tu, and S. Lazebnik, "Training deeper convolutional networks with deep supervision," *arXiv preprint arXiv:1505.02496*, 2015.

[25] L. Wang, X. Li, Y. Liao, Z. Jiang, J. Wu, F. Wang, C. Qian, and S. Liu, "Head: Hetero-assists distillation for heterogeneous object detectors," in *European Conference on Computer Vision*, pp. 314–331, Springer, 2022.

[26] L. Yao, R. Pi, H. Xu, W. Zhang, Z. Li, and T. Zhang, "G-detkd: Towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3591–3600, 2021.

[27] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *Advances in neural information processing systems*, vol. 35, pp. 38571–38584, 2022.