HamiPose: Hamiltonian Optimization for Unsupervised Domain Adaptive Pose Estimation

Supplementary Material

Table 6. Key notations.

Symbol	Meaning
$\mathcal{S}$	Labeled source domain dataset
$\mathcal{T}$	Unlabeled target domain dataset
x^s, y^s	Source image and its annotation
x^T	Target image (unlabeled)
$\mathbf{H}$	Predicted heatmaps
$f_\theta(\cdot)$	Student pose estimator
$f_{\theta'}(\cdot)$	Teacher pose estimator
τ	EMA decay for updating the teacher
$\mathcal{L}_s, \mathcal{L}_t, \mathcal{L}$	Source loss, target consistency loss, and total loss
$D_k^{(t-1)}$	Group-wise metric for keypoint channel k
λ	Weight of the target consistency term
g_s^{out}, g_t^{out}	gradients for source / target objectives
$g_{\text{total}}^{\text{out}}$	Filtered combined output space gradient
$\alpha^{(k)}$	Group-wise second-moment accumulator
μ	EMA decay for $\alpha^{(k)}$
ε	Small constant for numerical stability
p	Momentum variable in Hamiltonian optimization
ϵ	Step size of the symplectic update

A. Parameter-space alignment and projection

We decompose the final-layer parameters by keypoint/channel, writing the set responsible for the k-th heatmap as

$$\theta^{(k)} = \{W_{k,:}, b_k\}.$$

For this parameter block, let $J_k = \partial H_k / \partial \theta^{(k)}$ be the Jacobian of the channel output. Given output-space gradients $g_s^{(k)}$ and $g_t^{(k)}$, the chain rule yields the corresponding parameter-space gradients:

$$\hat{g}_s^{(k)} = J_k^\top g_s^{(k)}, \qquad \hat{g}_t^{(k)} = J_k^\top g_t^{(k)}. \tag{21}$$

These quantities measure how each channel influences the parameters.

To capture the geometry of updates applied throughout training, each block $\theta^{(k)}$ is equipped with a diagonal metric

$$D_k^{(t-1)} = (\alpha_{t-1}^{(k)} + \varepsilon)\, I, \qquad \varepsilon > 0, \tag{22}$$

where the scalar accumulator tracks recent squared update magnitudes:

$$\alpha_{t-1}^{(k)} = \mu\, \alpha_{t-2}^{(k)} + (1-\mu)\, \text{mean}\Big((\Delta\theta_{t-1}^{(k)})^2\Big), \qquad \alpha_0^{(k)} = 1. \tag{23}$$

This metric effectively rescales each block according to the curvature implied by its past motion.

Using this metric, we define the cosine-like alignment between source and target gradients:

$$\rho_k = \frac{(\hat{g}_s^{(k)})^\top D_k^{-1} \hat{g}_t^{(k)}}{\sqrt{(\hat{g}_s^{(k)})^\top D_k^{-1} \hat{g}_s^{(k)}}\, \sqrt{(\hat{g}_t^{(k)})^\top D_k^{-1} \hat{g}_t^{(k)}}}. \tag{24}$$

This quantity captures whether the two domains produce mutually reinforcing or conflicting signals under the metric.

The corresponding projection amplitude is

$$a_k = \frac{(\hat{g}_s^{(k)})^\top D_k^{-1} \hat{g}_t^{(k)}}{(\hat{g}_s^{(k)})^\top D_k^{-1} \hat{g}_s^{(k)}}. \tag{25}$$

This coefficient measures how much of the target gradient aligns with the source direction.

Although the coefficients are computed in parameter space, the projection itself is performed in output space:

$$\begin{aligned} g_{t,\parallel}^{(k)} &= a_k\, g_s^{(k)}, \\ g_{t,\perp}^{(k)} &= g_t^{(k)} - g_{t,\parallel}^{(k)}. \end{aligned} \tag{26}$$

This separation ensures that only the aligned component contributes to cross-domain transfer.

B. Non-conflict property under the parameter-space metric

We now show that the filtered update is non-conflicting with the source gradient after pullback. Let

$$\widehat{g}_{t,\perp}^{(k)} = J_k^\top g_{t,\perp}^{(k)}, \qquad \widehat{g}_s^{(k)} = J_k^\top g_s^{(k)}.$$

By substituting (26) and (25) into the definition,

$$\widehat{g}_{t,\perp}^{(k)} = \hat{g}_t^{(k)} - \frac{(\hat{g}_s^{(k)})^\top D_k^{-1} \hat{g}_t^{(k)}}{(\hat{g}_s^{(k)})^\top D_k^{-1} \hat{g}_s^{(k)}}\, \hat{g}_s^{(k)}. \tag{27}$$

This expression explicitly subtracts from $\hat{g}_t^{(k)}$ its projection onto the source direction.

Taking the metric inner product with $\widehat{g}_s^{(k)}$ gives

$$\begin{aligned} (\widehat{g}_{t,\perp}^{(k)})^\top D_k^{-1} \widehat{g}_s^{(k)} &= (\hat{g}_t^{(k)})^\top D_k^{-1} \hat{g}_s^{(k)} \\ &\quad - \frac{(\hat{g}_s^{(k)})^\top D_k^{-1} \hat{g}_t^{(k)}}{(\hat{g}_s^{(k)})^\top D_k^{-1} \hat{g}_s^{(k)}} (\hat{g}_s^{(k)})^\top D_k^{-1} \hat{g}_s^{(k)} \\ &= 0. \end{aligned} \tag{28}$$

Thus the orthogonal component remains orthogonal under the parameter-space metric after pullback.

To incorporate teacher confidence and cross-domain agreement, the filtered target gradient is

$$\tilde{g}^{(k)}_{t,\mathrm{pc}} = g^{(k)}_{t,\perp} + \phi_k\, a_k\, g^{(k)}_s, \qquad \phi_k = \max\{0, \tanh(\gamma(t)\rho_k)\}\, c_k.$$

Pulling back,

$$\widehat{\tilde{g}}^{(k)}_{t,\mathrm{pc}} = \widehat{g}^{(k)}_{t,\perp} + \phi_k a_k \widehat{g}^{(k)}_s.$$

The metric inner product of the composed update with the source direction becomes

$$\begin{aligned}(\widehat{g}^{(k)}_s)^\top D_k^{-1} \widehat{\tilde{g}}^{(k)}_{t,\mathrm{pc}} &= (\widehat{g}^{(k)}_s)^\top D_k^{-1} \widehat{g}^{(k)}_{t,\perp} \\ &\quad + \phi_k a_k (\widehat{g}^{(k)}_s)^\top D_k^{-1} \widehat{g}^{(k)}_s \\ &= 0 + \phi_k a_k\, (\widehat{g}^{(k)}_s)^\top D_k^{-1} \widehat{g}^{(k)}_s \;\geq\; 0,\end{aligned} \tag{29}$$

since $\phi_k \geq 0$, $a_k > 0$ whenever $\rho_k > 0$, and the quadratic form is positive. Hence the composed update is non-conflicting with respect to the source direction under the parameter-space metric.

C. Datasets Details

We employ three human pose datasets and three hand pose datasets that span both synthetic and real image domains. These datasets provide diverse appearance variations, annotation styles, and environmental conditions, enabling a comprehensive evaluation of cross-domain generalization.

Human pose datasets. For human pose estimation, we use SURREAL [34] as the synthetic source domain. SURREAL is rendered from SMPL body models under randomized textures, illuminations, and motions, and contains over six million images with dense 2D pose and segmentation annotations. Its large-scale and controlled rendering environment make it suitable for generating a wide range of synthetic poses, though its synthetic appearance introduces significant domain gaps.

As real-world target domains, we consider Human3.6M [13] and Leeds Sports Pose (LSP) [17]. Human3.6M consists of 3.6 million high-resolution frames captured from multiple synchronized cameras in an indoor motion-capture environment. Despite its controlled setting, it exhibits realistic lighting and clothing variations and is widely used for evaluating real-world human pose models. Following standard protocols [14, 18], we use subjects S1, S5, S6, S7, and S8 for training, and reserve S9 and S11 for testing to ensure a clean subject-level split. LSP contains 2,000 images collected from sports scenes in unconstrained outdoor environments. It includes challenging poses, self-occlusion, foreshortening, and diverse backgrounds. All images are used as unlabeled target samples during adaptation due to the relatively small dataset size.

Method	SURREAL→H36M		RHD→H3D	
	All	Improvement	All	Improvement
No-mom.	78.5	–	82.3	–
Mom.	79.1	+0.6	82.6	+0.3
HT (Ours)	**79.8**	**+1.3**	**83.1**	**+0.8**

Module	SURREAL→H36M		RHD→H3D	
	MeanCos↑	NegCos↓	MeanCos↑	NegCos↓
Backbone	0.24	0.19	0.22	0.18
Head	**0.12**	**0.34**	**0.11**	**0.32**

Method	SURREAL→H36M			RHD→H3D		
	Early↑	NegCos↓	Final↑	Early↑	NegCos↓	Final↑
None	70.0	0.41	77.2	75.5	0.38	80.9
Orth	75.8	0.21	77.6	79.0	0.19	81.5
Full	75.2	0.26	78.5	78.6	0.23	82.3

Table 7. No-mom.: SGD without momentum. Mom.: SGD with momentum. HT: Hamiltonian Transport. MeanCos/NegCos: mean and negative rate of $\cos(g_s, g_t)$, computed online at 10/50/90% training progress and averaged over 100 training iterations per phase. Early: performance at 10% of the training process.

Hand pose datasets. For hand pose estimation, we adopt the Rendered Hand Pose Dataset (RHD) [42] as the synthetic source domain. RHD contains 43,986 rendered images with accurate hand annotations. We follow the official split and use 41,258 images for training and 2,728 for validation. The rendering pipeline introduces a clear synthetic-to-real gap, especially in texture realism and hand–object interactions.

As real-world target domains, we select Hand-3D-Studio (H3D) [40] and FreiHand [43]. H3D provides 22,000 real RGB frames with 3D and 2D hand annotations captured using a multi-view camera rig. Following common practice [14, 18], we use 18,800 frames for training and the remaining for testing. The dataset features diverse lighting conditions and realistic articulation, making it a strong benchmark for real-world adaptation. FreiHand contains 130,000 real images covering a wide spectrum of poses, viewpoints, and hand–object interactions. All images are used as target-domain samples. Compared with H3D, FreiHand exhibits more complex hand appearance changes, stronger occlusion patterns, and higher variability in camera viewpoints.

In addition, we include the in-the-wild COCO Whole-Body Hand Dataset (WBH). WBH is generated by cropping hand instances from the COCO Whole-Body dataset [16], resulting in 76K training samples and 3.8K testing samples, each annotated with 21 keypoints. WBH features diverse backgrounds, spontaneous activities, and significant illumination variations, providing a real-world distribution that differs substantially from both RHD and FreiHand.

D. More Ablation Study

(1) Hamiltonian Transport vs. Plain Momentum. Table 7 shows that plain momentum brings only limited improve-

ment over vanilla SGD, whereas our Hamiltonian Transport (HT) achieves consistently larger gains on both transfer settings. This indicates that the advantage of HT does not simply come from adding inertia to the optimization. Instead, HT reshapes the target-driven update based on its relation to the source gradient, so that conflicting target signals are better controlled while useful adaptation cues can still be preserved. As a result, HT provides a more effective optimization behavior than standard momentum under noisy pseudo supervision.

(2) Layer-wise Analysis. As shown in Table 7, the prediction head consistently exhibits the lowest MeanCos and the highest NegCos, indicating that pseudo-label-induced gradient conflict is mainly concentrated in the final prediction layers. This is expected because the head directly maps intermediate representations to heatmaps and is therefore more sensitive to pseudo-label noise. In contrast, the backbone learns more general visual representations and is less directly affected by erroneous target supervision. This observation supports our design choice of refining only the head: it focuses on the most conflict-prone part of the network, reduces the risk of noisy gradients propagating into the backbone, and avoids the extra cost of full-network refinement.

(3) Orthogonal-only vs. Gated-parallel. Table 7 further isolates the roles of the two components. The *Orth* variant uses $g_{\text{upd}} = g_s + (g_t - a g_s)$, which removes the source-aligned component from the target gradient and keeps only the orthogonal residual. This effectively reduces gradient conflict and improves early-stage performance, showing that suppressing conflicting target directions is particularly helpful when pseudo labels are still noisy. However, the orthogonal-only design may also discard target information that is already beneficial and aligned with the source objective. The *Full* version therefore further introduces a confidence- and alignment-gated parallel term, which selectively restores useful aligned target signals. Consequently, *Orth* mainly improves optimization stability and early convergence, while *Full* achieves better final adaptation performance by balancing conflict suppression and target signal preservation.

E. Extended hyper-parameter study

Table 8 reports the ablation studies of alignment schedule in Eq. (10). We vary the initial sharpness $\gamma_{\min}$, the final sharpness $\gamma_{\max}$, and the warmup ratio r_{warm} and measure target PCK.

The rows with $\gamma_{\min} = 0$ consistently achieve the highest scores. Increasing $\gamma_{\min}$ to 0.5 or 1.0 slightly degrades performance but does not change the overall trend. This supports the choice of starting from a nonsharpened regime. For a fixed warmup ratio, performance peaks at a moderate $\gamma_{\max}$. The best results appear around $\gamma_{\max} = 2$, while smaller values reduce the benefit of aligned target gradients and larger values do not provide further gains. Across $\gamma_{\max}$ the differences remain within a narrow band, which indicates low sensitivity. Varying the warmup ratio shows that $r_{\text{warm}} = 0.10$ gives the most reliable improvement across different $\gamma_{\min}$ and $\gamma_{\max}$. Shorter warmup ratios use the sharp gate too early and slightly lower the final PCK, whereas longer warmup ratios delay adaptation. Overall, the table reveals a broad plateau around $\gamma_{\min} = 0$, $\gamma_{\max} = 2$, and $r_{\text{warm}} = 0.10$, which matches the default setting used in the main experiments.

Table 8. Ablation of the alignment schedule in Eq. 10. Reported values are target PCK scores (%) on RHD → H3D. Best result is highlighted in bold.

$\gamma_{\min}$	Warmup Ratio	$\gamma_{\max}$				
		1.0	1.5	2.0	2.5	3.0
0.0	0.05	68.0	68.8	70.1	69.7	69.3
0.0	0.10	68.2	69.2	**70.3**	70.0	69.6
0.0	0.15	68.0	69.0	70.0	69.6	69.2
0.5	0.05	67.6	68.4	69.4	69.0	68.8
0.5	0.10	67.8	68.9	69.7	69.2	69.0
0.5	0.15	67.6	68.6	69.3	69.0	68.7
1.0	0.05	67.3	68.1	68.6	68.3	68.0
1.0	0.10	67.5	68.3	69.0	68.7	68.4
1.0	0.15	67.4	68.2	68.7	68.5	68.2

These observations show that the proposed alignment schedule admits a wide range of workable configurations. The chosen default lies near the centre of a stable region rather than at an isolated optimum, which supports the robustness of our method with respect to alignment sharpness hyper-parameters.

F. Qualitative Comparisons

Figures 6 and 7 provide qualitative comparisons on human and hand pose estimation under cross-domhahaain settings. Compared with PCDA and DA-LLPose, our method produces predictions that are consistently closer to the ground truth, especially in challenging cases involving occlusion, illumination variation, truncation, and large viewpoint changes. For human pose estimation, competing methods often exhibit noticeable joint drift or structurally inconsistent predictions when the target appearance differs significantly from the source domain, whereas our method preserves a more coherent body configuration and yields more accurate localization of hard joints such as wrists, elbows, and ankles. For hand pose estimation, the advantage is even more evident in fingertip localization and articulated finger structure: baseline methods are more likely to produce distorted finger layouts or inaccurate tip positions, while our predictions better respect the underlying hand geometry and remain more stable under severe pose deformation and appearance changes.

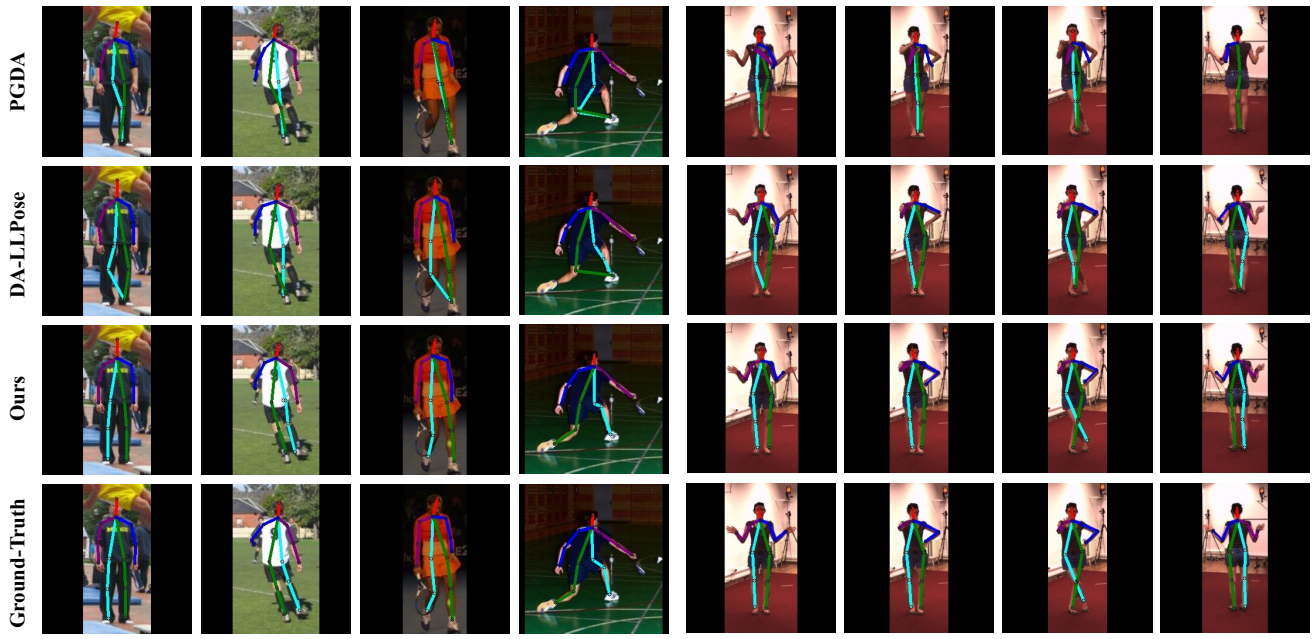

Figure 6. Qualitative comparison on human pose estimation. We compare PCDA, DA-LibPose, and our method on challenging cross-domain cases from Human3.6M and LSP. Each column shows a different test instance, and each row corresponds to a different method. Our approach produces pose predictions that more closely follow the ground-truth structure, especially under occlusion, illumination changes, and truncation.

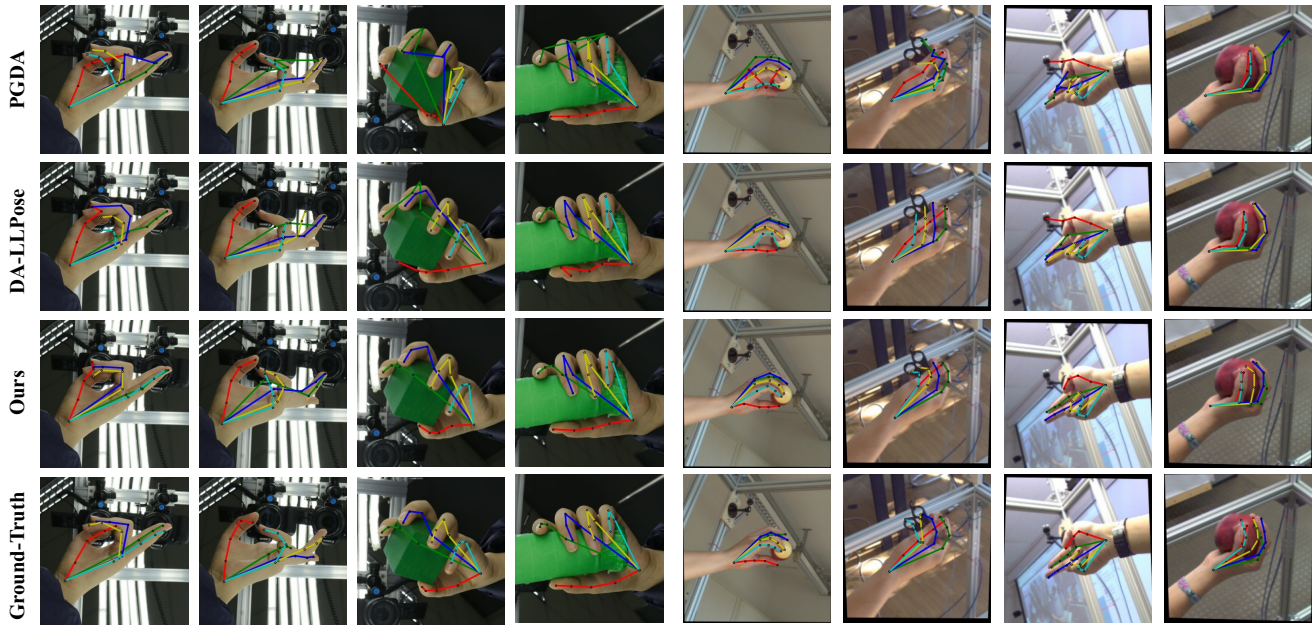

Figure 7. Qualitative comparison on hand pose estimation. Visualizations on cross-domain hand pose benchmarks comparing PCDA, DA-LibPose, and our method. Rows show predictions from different methods, and columns show diverse test instances with viewpoint and articulation variation. Our method achieves more accurate fingertip localization and overall hand structure alignment, producing results closest to the ground truth.

References

[1] Yihao Ai, Yifei Qi, Bo Wang, Yu Cheng, Xinchao Wang, and Robby T Tan. Domain-adaptive 2d human pose estimation via dual teachers in extremely low-light conditions. In *Proceedings of the European Conference on Computer Vision*, pages 221–239. Springer, 2024. 2, 6, 7

[2] Roberto Alcover-Couso, Marcos Escudero-Viñolo, Juan C SanMiguel, and Jesus Bescos. Gradient-based class weighting for unsupervised domain adaptation in dense prediction visual tasks. *Pattern Recognition*, 166:111633, 2025. 3

[3] Aristotelis Ballas and Christos Diou. Gradient-guided annealing for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20558–20568, 2025. 3, 7

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12397–12406, 2021. 1

[5] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J. Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6607–6616, 2020. 1

[6] Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, and Ke Lu. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3936–3945, 2021. 1, 2, 3, 6, 7

[7] Arindam Dutta, Sarosij Bose, Saketh Bachu, Calvin-Khang Ta, Konstantinos Karydis, and Amit K Roy-Chowdhury. Unsupervised domain adaptation for occlusion resilient human pose estimation. *arXiv preprint arXiv:2501.02773*, 2025. 2

[8] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. 2, 3, 7

[9] Zhiqiang Gao, Shufei Zhang, Kaizhu Huang, Qiufeng Wang, and Chaoliang Zhong. Gradient distribution alignment certificates better adversarial domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8917–8926, 2021. 1, 3, 6, 7

[10] Zhongyi Han, Haoliang Sun, and Yilong Yin. Learning transferable parameters for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 31:6424–6439, 2022. 2

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

[12] Yan He, Fei Peng, Rizhao Cai, Zitong Yu, Min Long, and Kwok-Yan Lam. Category-conditional gradient alignment for domain adaptive face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 19:10071–10085, 2024. 3

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 5, 2

[14] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6780–6789, 2021. 1, 2, 5, 6, 7

[15] Rui Jin, Jing Zhang, Jianyu Yang, and Dacheng Tao. Multi-branch adversarial regression for domain adaptative hand pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6125–6136, 2022. 2

[16] 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*, pages 196–214, 2020. 7, 2

[17] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, page 5, 2010. 5, 2

[18] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 603–620, 2022. 1, 2, 6, 7

[19] Youngho Kim, Hoonhee Cho, and Kuk-Jin Yoon. From sharp to blur: Unsupervised domain adaptation for 2d human pose estimation under extreme motion blur using event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9406–9417, 2025. 2

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. 8

[21] Binh M. Le and Simon S. Woo. Gradient alignment for cross-domain face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 188–199, 2024. 3

[22] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1491, 2021. 2, 6, 7

[23] Lijun Li, Linrui Tian, Xindi Zhang, Qi Wang, Bang Zhang, Liefeng Bo, Mengyuan Liu, and Chen Chen. Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20338–20348, 2023. 1

[24] Qiuxia Lin, Linlin Yang, and Angela Yao. Cross-domain 3d hand pose estimation with dual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17184–17193, 2023. 1

[25] Lucas Mansilla, Rodrigo Echeveste, Diego H Milone, and Enzo Ferrante. Domain generalization via gradient surgery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6610–6618, 2021. 1, 2, 3

[26] Jiteng Mu, Weichao Qiu, Gregory D. Hager, and Alan L. Yuille. Learning from synthetic animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12386–12395, 2020. 2, 6, 7

[27] Binh Nguyen, Minh-Duong Nguyen, Jinsun Park, Viet Pham, and Won-Joo Hwang. Federated domain generalization with data-free on-server gradient matching. In *International Conference on Learning Representations*, pages 77488–77513, 2025. 3

[28] Takehiko Ohkawa, Yu-Jhe Li, Qichen Fu, Ryosuke Furuta, Kris M Kitani, and Yoichi Sato. Domain adaptive hand keypoint and pixel localization in the wild. In *Proceedings of the European Conference on Computer Vision*, pages 68–87, 2022. 1

[29] Qucheng Peng, Ce Zheng, and Chen Chen. Source-free domain adaptive human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4803–4813, 2023. 1, 2, 6, 7

[30] Qucheng Peng, Ce Zheng, and Chen Chen. A dual-augmentor framework for domain generalization in 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2240–2249, 2024. 2

[31] Hoang Phan, Lam Tran, Quyen Tran, and Trung Le. Enhancing domain adaptation through prompt gradient alignment. *Advances in Neural Information Processing Systems*, 37:45518–45551, 2024. 1, 3, 6, 7

[32] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022. 3, 7

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pages 1195–1204, 2017. 2, 3

[34] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4627–4635, 2017. 5, 2

[35] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3769–3778, 2023. 3, 7

[36] Yikang Wei and Yahong Han. Multi-source collaborative gradient discrepancy minimization for federated domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15805–15813, 2024. 3

[37] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision*, pages 472–487, 2018. 5

[38] Yuanqi Yao, Gang Wu, Kui Jiang, Siao Liu, Jian Kuai, Xianming Liu, and Junjun Jiang. Improving domain generalization in self-supervised monocular depth estimation via stabilized adversarial training. In *Proceedings of the European Conference on Computer Vision*, pages 183–201, 2024. 3

[39] Ruipeng Zhang, Ziqing Fan, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Domain-inspired sharpness aware minimization under domain shifts. In *International Conference on Learning Representations*, pages 54751–54777, 2024. 3

[40] Zhengyi Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. Hand-3d-studio: A new multi-view system for 3d hand reconstruction. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2478–2482, 2020. 5, 2

[41] Zhan Zhuang, Yu Zhang, and Ying Wei. Gradual domain adaptation via gradient flow. In *International Conference on Learning Representations*, pages 26953–26978, 2024. 3, 6

[42] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4913–4921, 2017. 1, 5, 2

[43] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 5, 2