

Visibility guided Self-Supervised Occlusion-Resilient Human Pose Estimation

Arindam Dutta¹ Sarosij Bose¹ Rohit Kundu¹ Calvin-Khang Ta^{1,2,†} Saketh Bachu^{1,3,†}
 Konstantinos Karydis¹ Amit K. Roy-Chowdhury¹

¹University of California, Riverside, CA ²Dolby Laboratories, USA ³EvenUp, San Francisco, USA

{adutt020, sbose007, rkund006, cta003, sbach008, karydis, amitrc}@ucr.edu

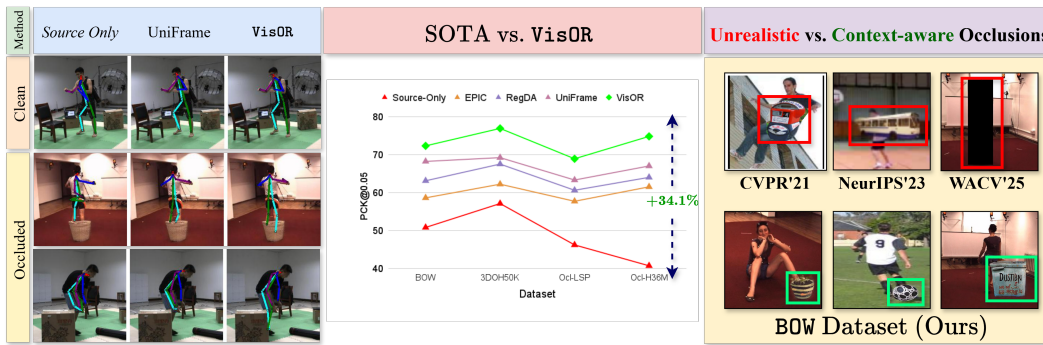


Figure 1. Overview. Left– **VisOR** significantly improves self-supervised 2D human pose estimation performance under occlusions. Compared against baselines, **VisOR** predicts accurate keypoint locations despite significant occlusions. Center– Quantitative results in four benchmark datasets (**BOW**, 3DOH50K, Occluded-LSP, Occluded-Human3.6M) show that **VisOR** outperforms existing domain adaptation methods [1, 2], particularly in challenging environments with high occlusions. Right– Comparison of occlusion types between datasets. Existing related works [3–5] often use **unrealistic semantically disconnected occluding** objects (e.g. black boxes or random textures shown in **red bounding boxes**). In contrast, our proposed **BOW** benchmark dataset introduces synthetic but **context-aware occlusions**, using objects relevant to the surrounding scene (shown in **green bounding boxes**). This leads to more natural occlusion patterns and enables better generalization to real-world occluded scenarios.

Abstract

Occlusion remains a significant challenge for existing human pose estimation algorithms, often resulting in inaccurate and anatomically implausible predictions. Although recent occlusion-robust methods report strong performance, they typically rely heavily on supervised learning and privileged information, such as multiview data or temporal sequences. Furthermore, these models often fail under domain changes. Domain-adaptive human pose estimation seeks to mitigate this issue; however, when occlusions are present in the target domain, a common occurrence in real-world applications, performance of these algorithms deteriorates significantly. To address these challenges, we propose **VisOR**, a novel **Visibility guided Self-Supervised** algorithm for **Occlusion-Resilient Human Pose Estimation**. **VisOR** achieves robustness to both domain shifts

and occlusions by integrating contextual reasoning with iterative pseudo-label refinement. It mitigates the overfitting to noisy labels from occluded regions via a visibility-driven curriculum learning strategy, which progressively introduces the model to increasingly occluded training samples. Additionally, **VisOR** is regularized by a learned human pose prior that maintains anatomical plausibility throughout the adaptation process. Recognizing the scarcity of human pose datasets with realistic occlusions, we introduce **BOW: Blended Occlusions in-the-Wild**, a rigorously constructed context-aware synthetic benchmark designed to evaluate the occlusion resilience of human pose estimation algorithms. **BOW** offers a diverse range of **context-aware occlusions** across both indoor and outdoor environments, simulating real-world conditions. Through extensive experiments, we demonstrate that **VisOR** outperforms current state-of-the-art methods by $\sim 7\%$ in challenging occluded human pose estimation benchmarks and provides a baseline performance on **BOW**, against existing algorithms.

[†] Work done while the authors were at UCR.

Project Webpage: <https://github.com/arid29/VisOR>

1. Introduction

Estimation of human pose is an important problem that aims to accurately detect anatomical keypoints in the human body. Precise pose estimates are particularly crucial for multimedia applications [6–8]. The remarkable success of state-of-the-art human pose estimation algorithms [9–11] in existing datasets [12, 13], acquired in controlled settings, can be largely attributed to their access to extensive labeled training data. However, occlusions, a common phenomenon in real-life scenarios, pose a major bottleneck for such algorithms, often leading to incorrect and anatomically implausible pose estimates (as shown in Figure 1). These algorithms show notable performance drops when tested on human images with occlusions (see Figures 1 and 4). Furthermore, acquiring large amounts of annotated data in the real world is costly and time-consuming, and it can raise privacy concerns [14]. One direct way to address these problems is by training models on large amounts of synthetic data and then applying these models to real-world datasets. However, models trained on synthetic data often suffer performance degradation when evaluated on real-world data due to the inherent domain differences between the two distributions, as shown in [1, 2]. Moreover, no open-source synthetic datasets provide human images with occlusions, further limiting the generalizability of existing models when tested on in-the-wild images of humans with occlusions.

Using unsupervised domain adaptation (UDA), researchers have successfully transferred knowledge from labeled source datasets to unlabeled target datasets, addressing various computer vision tasks [15–21]. Specifically for pose estimation, recent studies [1, 2, 22] leveraged UDA to mitigate the suboptimal generalization of models trained on synthetic datasets under domain shifts while simultaneously reducing reliance on ground truth annotations. However, these algorithms face notable challenges when dealing with occlusions in images of the unlabeled target domain. On the other hand, algorithms like those proposed in [23–25] that address occlusion-based performance degradation are highly dependent on in-domain supervised learning. For example, [24] uses multiple frames along with their corresponding ground truth poses to incorporate spatio-temporal continuity to manage self-occlusions in humans, and [23] requires paired occluded and unoccluded images along with ground truth pose keypoints to train a Siamese network to handle occlusions. However, these algorithms cannot handle domain shifts [1, 2]. Thus, our goal is to adapt a human pose estimation model trained on labeled source data to an unlabeled target dataset, where human subjects are partially blocked by occlusions from inanimate objects. To address this challenge, we propose **VisOR**: Visibility guided Self-Supervised Occlusion Resilient Human pose Estimation.

VisOR is a novel context-aware self-supervised algorithm that learns to generalize in an unlabeled target domain using confident pseudo-labels for self-training, employing the Mean-Teacher framework [26]. However, we observe that trivially learning from the confident pseudo-labels of the target domain

is largely unreliable, as the model may still generate implausible poses. To this end, we introduce a novel visibility-based curriculum learning strategy that compels the model to train on more visible samples with fewer occlusions and then progressively shifts to less visible samples with higher occlusions. As occlusions are random in both nature and location, this approach allows the network to learn from more reliable pseudo-labels in the early stages of adaptation, thus preventing error accumulation that typically results from learning from incorrect pseudo-labels. Unlike existing works like DMH-CL [27], which requires GT-derived hardness and supervised losses, **VisOR** does not use any target-domain annotation—using silhouette masks as the sole proxy for visibility. Also, to inject anatomical context into unsupervised adaptation, we develop a human-pose prior whose zero-level set approximates the human-kinematics manifold. This prior regularizes self-training by penalizing anatomically implausible predictions and, unlike exemplar similarity, evaluates a plausibility energy: valid poses receive low energy, invalid ones high. Under heavy occlusion—when noisy pseudo-labels can push the model toward unrealistic configurations—this prior acts as a strong safeguard, suppressing anatomically impossible poses during adaptation. **VisOR** works out-of-the-box for in-the-wild images, without requiring prior knowledge of the unlabeled target domain images. Overall, we note that **VisOR** significantly outperforms state-of-the-art unsupervised domain adaptive human pose estimation algorithms when unlabeled target data are occluded.

Furthermore, recognizing the scarcity of human pose datasets with realistic occlusions, we introduce **BOW**: Blended Occlusions in-the-Wild, a rigorously constructed context-aware synthetic benchmark designed to evaluate the occlusion resilience of human pose estimation algorithms. Unlike existing synthetic datasets that randomly overlay objects without considering scene context, **BOW** employs a multimodal large language model (MLLM) [28] to generate context-aware occlusions by predicting plausible inanimate objects, such as “a rock,” “a shrub,” or “a basketball” (see Figures 1 and 3) that naturally fit within the scene. These objects are then synthesized and seamlessly composited into the original images using large-scale pretrained Text-to-Image models [29], ensuring photorealistic and semantically coherent occlusions. **BOW** introduces diverse objects occluding different parts of the human body in images in both indoor and outdoor settings, overcoming the limitations of previous datasets [30, 31] that lack semantic plausibility, as shown in Figure 1. **BOW** includes synthetic occlusion variants for standard benchmarks such as LSP [32] and Human3.6M [33], making it a valuable resource for developing robust pose estimation models capable of handling complex occlusion scenarios in real-world applications. *In this work, we use **BOW** to evaluate existing algorithms [1, 2, 22] and introduce a new baseline for realistic occlusions.*

Contributions: Our key contributions are as follows:

- We address the challenge of self-supervised, occlusion-

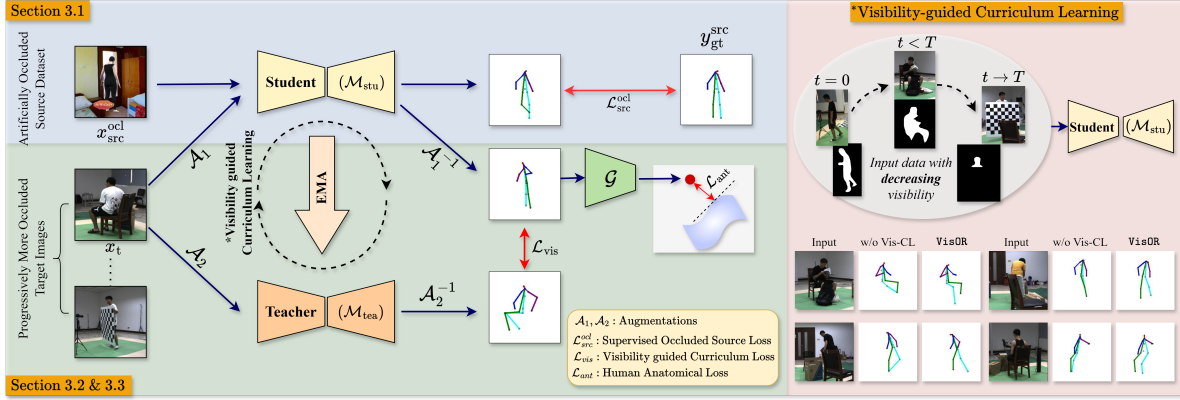


Figure 2. **Overview of proposed algorithm.** **VisOR** is built upon the mean-teacher framework, where the weights of the teacher model are updated as an exponential moving average (EMA) of the student model’s weights. **VisOR** uses occlusion augmentations on the source domain, enabling the student and teacher models to provide better pseudo labels on unlabeled target images by learning consistency between occluded and unoccluded source images. **VisOR** also utilizes a human pose prior that captures plausible human anatomy as a zero-level set, thus penalizing anatomically implausible predictions. Finally, **VisOR** leverages a visibility-based curriculum learning strategy wherein the model focuses on samples with fewer occlusions and gradually shifts to samples with higher degrees of occlusion.

resilient human pose estimation, specifically targeting the limitations of current unsupervised domain adaptation (UDA) methods under occlusion-heavy conditions.

- We propose **VisOR**, a self-training framework for occlusion-resilient human pose estimation in unlabeled target domains. **VisOR** combines a mean-teacher approach with iterative pseudo-label refinement and employs visibility-aware curriculum learning alongside a learned human pose prior to ensure anatomically plausible predictions.
- To mitigate the scarcity of occluded human datasets, we introduce **BOW**, a context-aware synthetic benchmark that generates semantically consistent occlusions using foundational generative models, facilitating rigorous and realistic evaluation of occlusion robustness.
- Extensive experiments across multiple challenging benchmarks—including **BOW**—demonstrate that **VisOR** achieves consistent improvements over state-of-the-art methods, outperforming them by $\sim 7\%$ on unlabeled occluded datasets.

2. Related Works

Human Pose Estimation: Human Pose Estimation involves localizing anatomical landmarks of the human body in images or videos. Large-scale annotated datasets have driven significant advances in 2D pose estimation [34]. Existing approaches are classified as top-down, which detects humans before estimating poses [35–38], and bottom-up, which detects keypoints first and then group them per person [9, 39–41]. These fully supervised models often fail to generalize to unseen domains and partial occlusions. To handle occlusions, [23] synthesized occluded views using siamese networks, while [24] used spatio-temporal continuity to transfer information from clear frames. However, these methods rely on paired clean and

occluded images or temporal consistency, which limits their applicability in real-world occluded settings. Recent methods include a multistage architecture for occluded multiperson scenes [25], which struggles with heavy occlusion, and a graph-based model [42] that infers missing keypoints but relies on strong spatial priors and fails on out-of-distribution poses. [43] introduced an occlusion-aware framework to infer visible 2D keypoints from partial input, but its reliance on labeled 2D data limits scalability to unlabeled or diverse real-world settings. Despite these advancements, most methods still depend on labeled data and struggle with domain shifts and occlusions.

Datasets for Pose Estimation Under Occlusions: Most human pose estimation benchmarks pay limited attention to occlusions, and collecting such data is challenging due to specialized hardware and annotation ambiguities. Prior works [4, 31] simulate occlusions via size-based masks or preprocessing; however, these methods are not context-aware and often ignore scene semantics, producing artifacts that models can exploit rather than learn true robustness. Recent datasets move closer to realistic occlusion scenarios but rely on different assumptions. Blend-Mimic3D [44] is a Blender-generated synthetic benchmark with labeled occlusion visibility; AGORA [45] renders photorealistic synthetic humans with complex occlusions; LiCamPose [46] targets multimodal RGB+LiDAR capture using both synthetic pretraining data and real recordings. In contrast, **BOW** retains real in-the-wild humans and synthesizes only the occluders with contextual awareness. This isolates the effect of occlusion without confounding changes in human appearance or capture hardware, yielding a more faithful test of occlusion robustness.

UDA for Human Pose Estimation: Several works have studied UDA for human pose estimation. [2, 22] introduced adversarial learning-based UDA frameworks, and UniFrame [1] employed a Mean-Teacher framework for pseudo-label refinement, yet

their performance is limited when the target domain contains occlusions, as they assume full-body visibility. [47] proposed a cross-domain UDA for animal poses, but anatomical differences limit its relevance to humans. However, these methods are trained and evaluated on occlusion-free benchmarks and do not explicitly address the challenge of partial occlusions, making them ill-suited for real-world applications. In contrast, we focus on the problem of 2D human pose estimation under occlusions in unsupervised domain adaptation settings. We show that the state-of-the-art UDA approaches (Table 1) perform suboptimally when human figures in the target domain are partially occluded, underscoring the need for self-supervised algorithms that are resistant to occlusion (**VisOR**).

3. Proposed Algorithm (**VisOR**)

In this section, we elaborate on our proposed algorithm (**VisOR**) for self-supervised occlusion-resilient human pose estimation. Our goal is to develop a human pose estimation model $\mathcal{F}$ that takes an image $x \in \mathcal{R}^{H \times W \times 3}$ with spatial dimensions H and W as input and predicts the 2D pose keypoints $y = \mathcal{F}(x)$, where $y \in \mathcal{R}^{K \times 2}$ for K keypoints. We assume access to a source dataset $\mathcal{S} = \{x_i, y_i\}_{i=1}^{N_1}$, where x_i is an RGB image and y_i is the corresponding set of pose keypoints. We also have an unlabeled target dataset $\mathcal{T}$, containing images $\{x_i\}_{i=1}^{N_2}$ without pose keypoint annotations. For clarity, we refer to an image from the labeled source dataset as x_s , with y_s being the corresponding set of pose keypoints, and an image from the unlabeled target set as x_t . We aim to adapt a model trained on the source dataset, denoted as $\mathcal{F}_S$, to enhance its performance on the target dataset $\mathcal{T}$. The model under adaptation/adapted model is denoted as $\mathcal{F}_T$.

We propose an intuitive self-learning algorithm for optimal human pose estimation in unsupervised domain adaptive settings, with a particular focus on situations where the human subject in the unlabeled target data is partially occluded by inanimate objects. Existing analogous algorithms for human pose estimation [1, 2, 22] often suffer from error accumulation due to learning incorrect pseudo-labels and a lack of contextual understanding of the human body. This results in suboptimal performance when the human subject is partially occluded in the unlabeled target data.

To address these, we introduce three distinct methods:

- We employ an exponentially weight-averaged teacher model to generate reliable pseudo-labels for self-training. This approach ensures that the teacher model does not overfit incorrect pseudo-labels. Additionally, we artificially occlude the labeled source images; this acts as an augmentation step thereby allowing the student model to learn estimating human poses under occlusions.
- A human pose prior, designed as a zero-level set for plausible human poses, is utilized to regularize the adaptation process by penalizing the model for predicting anatomically implausible pose predictions.
- Recognizing that not all unlabeled target images are equally

occluded, we enable the network under adaptation to initially focus on samples with less relatively occlusions. As the adaptation progresses, the network gradually addresses more heavily occluded samples. This approach prevents the network from overfitting to incorrect pseudo-labels from the more heavily occluded samples.

We frame pose estimation as a heatmap regression task [48], where the model is trained to predict Gaussian heatmaps centered on each keypoint location. Specifically, instead of directly regressing keypoint coordinates, the network learns to generate probability distributions that peak at ground-truth joint positions, effectively converting pose estimation into a structured spatial regression problem. The weights of $\mathcal{F}_T$ are initialized using the weights of a model that has been trained on labeled data from the source domain through supervised learning [1, 2, 22]. This model, known as the source model ($\mathcal{F}_S$), is trained by minimizing the loss function defined in Eqn. 1.

$$\mathcal{L}_{\text{src}} = \frac{1}{N_1} \sum_{x_s} \|\mathcal{F}_S(x_s) - y_s\|_2^2. \quad (1)$$

3.1. Self-Training and Occlusion Augmentations

We take advantage of the Mean-Teacher framework [26] for our problem settings. In the Mean-Teacher framework, two identical models are utilized: $\mathcal{M}_{stu}$ (student) and $\mathcal{M}_{tea}$ (teacher), with parameters denoted as θ_{stu} and θ_{tea} , respectively. Both models are initialized with the weights from a pretrained source model ($\mathcal{F}_S$). During each adaptation step t , the parameters of the student model (θ_{stu}) are updated by backpropagating the gradient computed in that step, while the parameters of the teacher model (θ_{tea}) are adjusted as an exponential moving average of the weights of the student model (θ_{stu}) and as a function of the smoothing parameter α as described in Eqn. 2. This approach aims to improve overall model performance by reducing overfitting to potentially incorrect pseudo-labels, particularly during the initial phase of adaptation.

$$\theta_{tea}^t \leftarrow \alpha \theta_{tea}^{t-1} + (1 - \alpha) \theta_{stu}^t \quad (2)$$

We enforce a prediction space consistency for unlabeled target images across two different augmentations. For any sampled target image x_t , we apply augmentations $\mathcal{A}_1(\cdot)$ and $\mathcal{A}_2(\cdot)$ to the same, resulting in augmented images $\mathcal{A}_1(x_t)$ and $\mathcal{A}_2(x_t)$. The corresponding heatmap predictions are obtained as $\mathcal{H}_1 = \mathcal{M}_{stu}(\mathcal{A}_1(x_t))$ and $\mathcal{H}_2 = \mathcal{M}_{tea}(\mathcal{A}_2(x_t))$. The pseudo labels are obtained by extracting the pixel coordinates with the highest activations from the heatmap predictions of the teacher model ($\mathcal{H}_2$). In order to avoid overfitting to potentially incorrect pseudo labels predicted by the teacher model, **VisOR** only learns from predictions that exceed a specified confidence threshold (τ) [1]. These confident pseudo-labels are then converted into normalized heatmaps, which are used to update $\mathcal{M}_{stu}$ by minimizing the loss function defined in Eqn. 3.

$$\mathcal{L}_{\text{pred}} = \frac{1}{|\mathcal{B}|} \sum_{x_t \in \mathcal{B}} \sum_{k \in K} \mathbf{1}(\mathcal{H}_2^k \geq \tau) \quad (3)$$

$$\|\mathcal{A}_1^{-1}(\mathcal{H}_1^k) - \mathcal{A}_2^{-1}(\mathcal{H}_2^k)\|_2^2.$$

Here, $\mathcal{A}_1^{-1}(\cdot)$ and $\mathcal{A}_2^{-1}(\cdot)$ are defined as inverse augmentations of $\mathcal{A}_1(\cdot)$ and $\mathcal{A}_2(\cdot)$, respectively, and $\mathcal{B}$ denotes a given batch of images. During this self-training process, we also artificially occluded the labeled source domain images (x_s) with objects from the COCO dataset [13], referring to these images as x_{src}^{ocl} . This allows the student model to generalize better to occlusions by leveraging the labeled source data and minimizing the loss function defined by Eqn. 4.

$$\mathcal{L}_{src}^{ocl} = \frac{1}{|\mathcal{B}|} \sum_{x_{src}^{ocl} \in \mathcal{B}} \|\mathcal{M}_{stu}(x_{src}^{ocl}) - y_s\|_2^2. \quad (4)$$

We observe that leveraging the loss functions described by Eqn. 3 and Eqn. 4 leads to improved performance over UniFrame [1]. However, the adapted model may still produce anatomically implausible poses due to a lack of human anatomical constraints under occlusions, thus leading to suboptimal overall performance.

3.2. Regularization with Human Pose Prior

To integrate anatomical constraints into the adaptation process, we introduce a parametric regularizer defined as the zero-level set of plausible human poses. Specifically, we represent 2D poses as connecting pairs of joints. Formally, a skeleton y is described by a set of vectors corresponding to these bone connections,

$$\theta_y = \{b_1, b_2, \dots, b_M\}; \quad b_m \in \mathcal{R}^2 \quad \forall m \in [M] \quad (5)$$

We propose to train the parametric human pose prior ($\mathcal{G}: \theta_y \rightarrow \mathcal{R}^+$) such that

$$\mathcal{G}(\theta_y) \begin{cases} = 0, & \theta_y \text{ is plausible;} \\ > 0, & \text{otherwise.} \end{cases} \quad (6)$$

We define $\mathcal{G} = \mathcal{G}_{dec} \circ \mathcal{G}_{enc}$, $\mathcal{G}_{enc}$ projecting individual vectors (b_m) to a latent space based on the anatomical structure of the human body [49, 50] and $\mathcal{G}_{dec}$ projects this feature in the latent space to a non-negative scalar representing the distance of the queried pose from the learned manifold.

Training $\mathcal{G}$ is performed using the labeled source dataset ($\mathcal{S}$), which contains N_1 accurate human poses. However, the source dataset ($\mathcal{S}$) does not include examples of implausible human poses, which are also important for learning the pose prior $\mathcal{G}$. To address this, we propose generating examples of implausible poses by artificially occluding the images in $\mathcal{S}$ ([51]) and estimating the pose using the $\mathcal{F}_S$ model, which is pretrained on clean, unoccluded images of $\mathcal{S}$.

For any image from the labeled source dataset (x_s), let y_s be the corresponding ground-truth pose label. Let $\hat{x}_s$ denote the image generated by artificially occluding x_s . The corresponding noisy pose estimates are obtained as $\hat{y}_s = \mathcal{F}_S(\hat{x}_s)$. We also generate a set of implausible poses by adding noise sampled from the Von-Mises distribution [52] to the individual keypoints.

We denote the distance between the plausible ground truth pose (y_s) and the correspondingly generated implausible pose ($\hat{y}_s$) as $\mathbf{d}$ which is set to 0 if θ_y is sampled from the distribution of poses encompassed by the source dataset $\mathcal{S}$, and $\mathbf{d} > 0$ otherwise. To train $\mathcal{G}$, we use supervised learning by minimizing the loss L_2 between the predicted score for a given pose θ_y and the corresponding ground truth score. Mathematically, this can be expressed as follows:

$$\underset{\mathcal{G}}{\text{argmin}} \quad \|\mathcal{G}(\theta_y) - \mathbf{d}\|_2^2 \quad (7)$$

During adaptation, **VisOR** leverages this trained human pose prior as a regularizer for the anatomical plausibility of the predicted poses on the unlabeled target domain images (x_t). Mathematically, we minimize:

$$\mathcal{L}_{ant} = \frac{1}{|\mathcal{B}|} \sum_{x_t \in \mathcal{B}} \mathcal{G}(\mathcal{M}_{stu}(x_t)). \quad (8)$$

3.3. Visibility guided Curriculum Learning

As real-world images often vary in their levels of occlusion, we hypothesize that adapting $\mathcal{F}_T$ using unlabeled target domain images with fewer occlusions initially and gradually progressing to images with more occlusions will enhance the performance of $\mathcal{F}_T$ by allowing it to learn from more robust and accurate pseudo-labels. Thus, we propose a novel curriculum learning strategy that incrementally adapts the student model ($\mathcal{M}_{stu}$) from images with fewer occlusions to those with relatively more occlusions.

To condition the adaptation process based on human visibility in occluded scenes, we propose using silhouette segmentation maps from the unlabeled target domain images. For given images in $\mathcal{T}$, we can extract binary silhouettes of the visible portions of humans using pretrained models such as [53, 54]. For all unlabeled target images from $\mathcal{T}$, we extract binary silhouettes $\{s_i\}_{i=1}^{|\mathcal{T}|}$ using an existing pretrained segmentation model ($\mathbf{S}$). Mathematically, $s_i = \mathbf{S}(x_i) \quad \forall x_i \in \mathcal{T}$. Using these extracted silhouettes $\{s_i\}_{i=1}^{|\mathcal{T}|}$, we define the visibility score (v_i) for a given image in batch ($\mathcal{B}$) sampled from $\mathcal{T}$ as: $v_i = \text{sum}(s_i) / \max(\text{sum}(s_i)) \quad \forall s_i \in \mathcal{B}$.

Here, the $\text{sum}(\cdot)$ operator is defined as the total count of foreground pixels in the visible silhouette. Clearly $v_i \in [0, 1]$, with a higher score reflecting that the sample is relatively less occluded. We use these visibility scores to guide the adaptation algorithm by weighing the contributions of individual images in a given batch ($\mathcal{B}$) sampled from $\mathcal{T}$. Thus, more visible images (hence, higher v_i) will have a larger contribution in $\mathcal{L}_{\text{pred}}^{\text{vis}}$ (Eqn. 9),

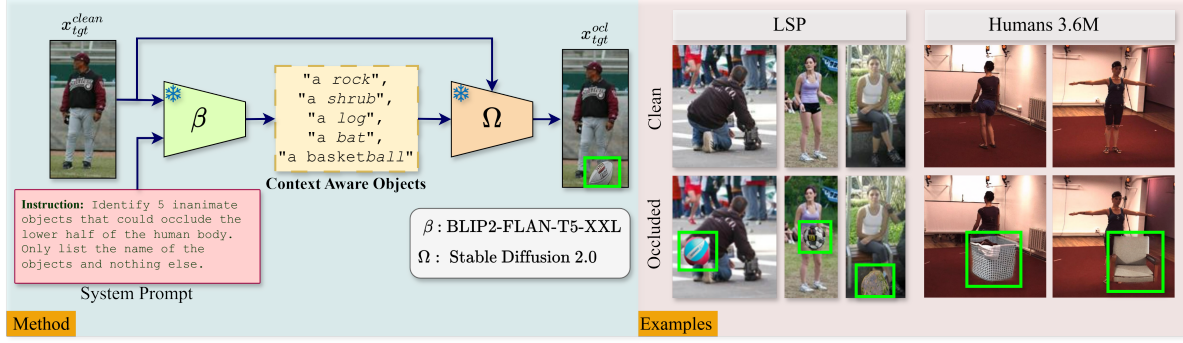


Figure 3. **Overview of BOW Benchmark Dataset Generation Pipeline.** For each clean image x_{tgt}^{clean} in the target domain, we first prompt an MLLM [28] β to generate a set of plausible object names which can blend in a mask placed within the target image. An LDM [29] (Ω) then blends in that object within the target image hence effectively placing occlusions leveraging the context of the visible portions of each image. We then combine these obtained blended images x_{tgt}^{occl} from the LSP [32] and Humans3.6M [33] datasets to create **BOW**.

thereby forcing the model to learn from those samples especially during the initial stages of adaptation. To this end, we propose:

$$\mathcal{L}_{pred}^{vis} = \frac{1}{|\mathcal{B}| \times \sum v_i} \sum_{x_t \in \mathcal{B}} \sum_{k \in K} \mathbf{1}(\mathcal{H}_2^k \geq \tau) \cdot v_i \|\mathcal{A}_1^{-1}(\mathcal{H}_1^k) - \mathcal{A}_2^{-1}(\mathcal{H}_2^k)\|_2^2. \quad (9)$$

Observe that using only the loss described in Eqn. 9 will largely ignore the heavily occluded images of $\mathcal{T}$. This is not desirable, as this could lead to suboptimal performance on heavily occluded images in $\mathcal{T}$. Thus, we resort to using a linear combination of the loss functions defined by Eqn. 9 and Eqn. 3. We define

$$\mathcal{L}_{vis} = \gamma \mathcal{L}_{pred}^{vis} + (1 - \gamma) \mathcal{L}_{pred}. \quad (10)$$

with $\gamma = \exp(\frac{-current\ epoch}{total\ epochs})$. At the beginning of adaptation, when $\gamma \rightarrow 1$, $\mathcal{L}_{vis}$ emphasizes samples with minimal occlusion; as adaptation progresses, the model gradually shifts to treating all samples in the batch more equally.

Final Objective: Leveraging the aforementioned loss functions, the student model ($\mathcal{M}_{stu}$) is trained using the weighted objective function, which is mathematically defined as:

$$\underset{\mathcal{M}_{stu}}{\operatorname{argmin}} \quad \mathcal{L}_{src}^{occl} + \lambda_a \mathcal{L}_{ant} + \lambda_v \mathcal{L}_{vis}. \quad (11)$$

4. Context Aware Occlusions in-the-Wild

Acknowledging the scarcity of human pose datasets with realistic occlusions, we introduce a context-aware synthetic benchmark (**BOW**) that enables systematic evaluation of occlusion robustness for human pose estimation algorithms. Our approach (Figure 3) generates semantically plausible occluded images conditioned on occlusion-free images from standard datasets such as LSP [32] and Human3.6M [33], using a two-stage generative pipeline that ensures photorealism and contextual relevance.

In the first stage, we pass the clean target image x_{tgt}^{clean} to a multimodal large language model (MLLM), denoted by

β [28], to generate names of plausible inanimate objects that could serve as occluding objects based on the image context. Specifically, we define a system prompt $\mathcal{P}$ comprising structured natural language instructions in a question-answer format. We employ a few-shot in-context learning setup [56], where the MLLM is shown example pairs $(x_{tgt}^{clean}, x_{tgt}^{occl})$ to guide it in producing semantically grounded and contextually appropriate object suggestions, avoiding unrealistic insertions seen in prior synthetic datasets [30, 51], as illustrated in Figure 1.

The MLLM outputs a list of candidate object tags $\mathcal{O}$, which we filter using token likelihoods to retain only the most contextually fitting objects for a given image-mask pair (x_{tgt}^{clean}, m) :

$$\tilde{\mathcal{O}} = \left\{ o_i \in \mathcal{O} \mid P_{\beta}(o_i \mid x_{tgt}^{clean}, m) \geq \tau \right\}. \quad (12)$$

We then select the top k object tags by likelihood:

$$\mathcal{O}_{topk} = \operatorname{TopK}_{o_i \in \tilde{\mathcal{O}}} \left(P_{\beta}(o_i \mid x_{tgt}^{clean}, m) \right), \quad (13)$$

where TopK returns the subset of size k with the highest scores from the filtered set $\tilde{\mathcal{O}}$.

In the second stage, we leverage these predicted object categories to guide a generative image synthesis model Ω , based on Stable Diffusion 2.0 [29, 57]. For each selected object $o \in \mathcal{O}_{topk}$, we define a binary spatial mask $m \in \{0, 1\}^{H \times W}$ that indicates the region where the object will be inserted. This mask is constructed by sampling a patch of size $h \times w$ centered around a randomly selected keypoint from a predefined set $\mathcal{K}$, ensuring that the synthesized object partially occludes the human figure.

Given an image $x_{tgt}^{clean} \in \mathbb{R}^{3 \times H \times W}$, we encode it into the latent space using the encoder $\mathcal{E}$ of the diffusion model:

$$z = \mathcal{E}(x_{tgt}^{clean}), \quad z \in \mathbb{R}^{C \times H' \times W'}. \quad (14)$$

During denoising, at each diffusion timestep t , we maintain two latent trajectories: the original latent z_t and an object-conditioned latent $\tilde{z}_t$, guided by a text prompt $\tau(o)$ describing

Performance of VisOR vs. existing baselines: PCK@0.05 across different joints. Best results highlighted in **bold** and gain over *Source only* is highlighted in **green**.

Algorithm	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg.
<i>Source only</i>	49.4	57.2	47.9	40.3	56.7	53.2	50.8
EPIC [22]	65.1	63.5	47.6	51.8	64.2	59.2	58.6
RegDA [2]	62.4	69.6	57.8	54.5	67.8	66.5	63.1
UniFrame [1]	62.5	75.2	66.0	63.6	72.9	68.8	68.2
VisOR	66.1	79.6	69.7	67.4	77.3	73.0	72.3
	(+16.7)	(+22.4)	(+21.8)	(+27.1)	(+20.6)	(+19.8)	(+21.5)

Table 1. SURREAL [51] → BOW

Algorithm	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg.
<i>Source only</i>	74.7	62.5	50.4	41.2	57.9	56.0	57.1
EPIC [22]	81.2	65.5	42.5	48.2	64.3	71.6	62.2
RegDA [2]	85.5	73.3	56.5	50.5	67.9	71.0	67.5
UniFrame [1]	55.3	75.5	68.2	68.0	78.9	69.4	69.2
VisOR	78.7	81.4	72.2	70.7	82.2	76.4	76.9
	(+4.0)	(+18.9)	(+21.8)	(+29.5)	(24.3)	(+20.4)	(+19.8)

Table 2. SURREAL [51] → 3DOH50K [55]

Algorithm	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg.
<i>Source only</i>	44.7	53.5	49.5	43.5	40.7	45.2	46.2
EPIC [22]	53.2	61.7	53.3	63.5	56.3	58.2	57.7
RegDA [2]	55.0	65.0	55.8	66.6	59.8	61.6	60.6
UniFrame [1]	58.0	72.1	65.3	68.3	55.2	60.7	63.3
VisOR	59.5	75.7	67.8	74.7	65.2	70.2	68.9
	(+14.8)	(+22.2)	(+18.3)	(+31.2)	(+24.5)	(+25.0)	(+22.7)

Table 3. SURREAL [51] → Ocl-LSP

Algorithm	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg.
<i>Source only</i>	49.1	49.3	37.0	26.1	41.3	41.1	40.7
EPIC [22]	71.9	67.3	52.3	48.1	63.7	65.4	61.5
RegDA [2]	68.7	73.1	53.7	52.7	66.7	69.1	64.0
UniFrame [1]	69.7	77.8	61.4	57.7	70.9	64.5	67.0
VisOR	73.2	85.0	68.4	67.8	79.2	75.0	74.8
	(+24.1)	(+35.7)	(+31.4)	(+41.7)	(+37.9)	(+33.9)	(+34.1)

Table 4. SURREAL [51] → Ocl-H36M

the occluding object. These are blended spatially to form:

$$z'_t = m \odot \tilde{z}_t + (1 - m) \odot z_t, \quad (15)$$

where m is resized to match the latent resolution (H', W'), and $\odot$ denotes element-wise multiplication.

The object-conditioned latent $\tilde{z}_t$ is computed using classifier-free guidance [58]:

$$\tilde{z}_t = \epsilon_\theta(z_t, t, \tau(o)) + w \cdot [\epsilon_\theta(z_t, t, \tau(o)) - \epsilon_\theta(z_t, t, \emptyset)], \quad (16)$$

where ϵ_θ is the noise prediction network, $\tau(o)$ is the object prompt, $\emptyset$ denotes an empty prompt, and w is the guidance weight controlling insertion strength. After denoising to obtain z'_0 , the final occluded image is generated via decoding:

$$x_{\text{tgt}}^{\text{ocl}} = \mathcal{D}(z'_0), \quad (17)$$

where $\mathcal{D}$ is the decoder of the diffusion model. This results in an image $\hat{x}$ in which the occluding object described by $\tau(o)$ is realistically synthesized in regions defined by m , while preserving the context of the original image.

Our proposed pipeline facilitates the generation of **BOW** comprising high-quality occluded image variants derived from occlusion-free datasets, covering indoor and outdoor scenarios. By ensuring semantic consistency and photorealistic rendering, the resulting benchmark serves as a rigorous and controlled testbed for assessing the occlusion robustness of human pose estimation models, while also being able to use corresponding occlusion-free ground-truth annotations for quantitative evaluations.

5. Experiments

In this section, we showcase **VisOR**'s capability to estimate human poses under occlusions across various unlabeled target datasets. Further implementation and dataset details, along with additional experiments and results, are provided in the supplementary material.

Quantitative Results: In Table 1, we report the performance of **VisOR** on the Surreal → **BOW** benchmark. This benchmark captures a diverse range of scenes that span both indoor and outdoor environments, **VisOR** outperforms all existing algorithms [1, 2, 22] by approximately 5%, demonstrating its strong generalization and effectiveness. Table 2 presents the results on the Surreal → 3DOH50K benchmark, where the target domain features naturally occurring occlusions in real-world settings. These occlusions are diverse and unstructured, arising from interactions between humans and objects in cluttered backgrounds. In this challenging scenario, **VisOR** achieves a significant performance gain ($\sim 7\%$) over existing algorithms. This result underscores the strength of **VisOR** in handling complex, real-world occlusions without supervised data in the target domain. Finally, we examine the performance of Surreal → Ocl-LSP and Surreal → Ocl-H36M, shown in Tables 3 and 4, respectively. These benchmarks introduce synthetic occlusions into the respective datasets, simulating challenging visibility conditions. Across both settings, **VisOR** maintains consistent improvements over baselines [1, 2, 22], demonstrating its robustness to occlusions.

Qualitative Results: We present qualitative results for all benchmarks, Surreal → **BOW**, Surreal → 3DOH50K, Surreal → Ocl-LSP, and Surreal → Ocl-H36M, in Figure 4. As illustrated, the best-performing existing method, UniFrame [1], struggles to estimate accurate human poses when occlusions are present in unlabeled target images. It often produces anatomically implausible predictions, such as the misplacement of the left ankle on the right side and occlusions that further exacerbate these errors. In contrast, **VisOR** consistently generates accurate and coherent poses across all benchmarks, highlighting its robustness and effectiveness in addressing occlusions within the unsupervised domain adaptation setting.

Ablation Study: In Table 5, we present an ablation study that analyzes the contribution of each component of our proposed

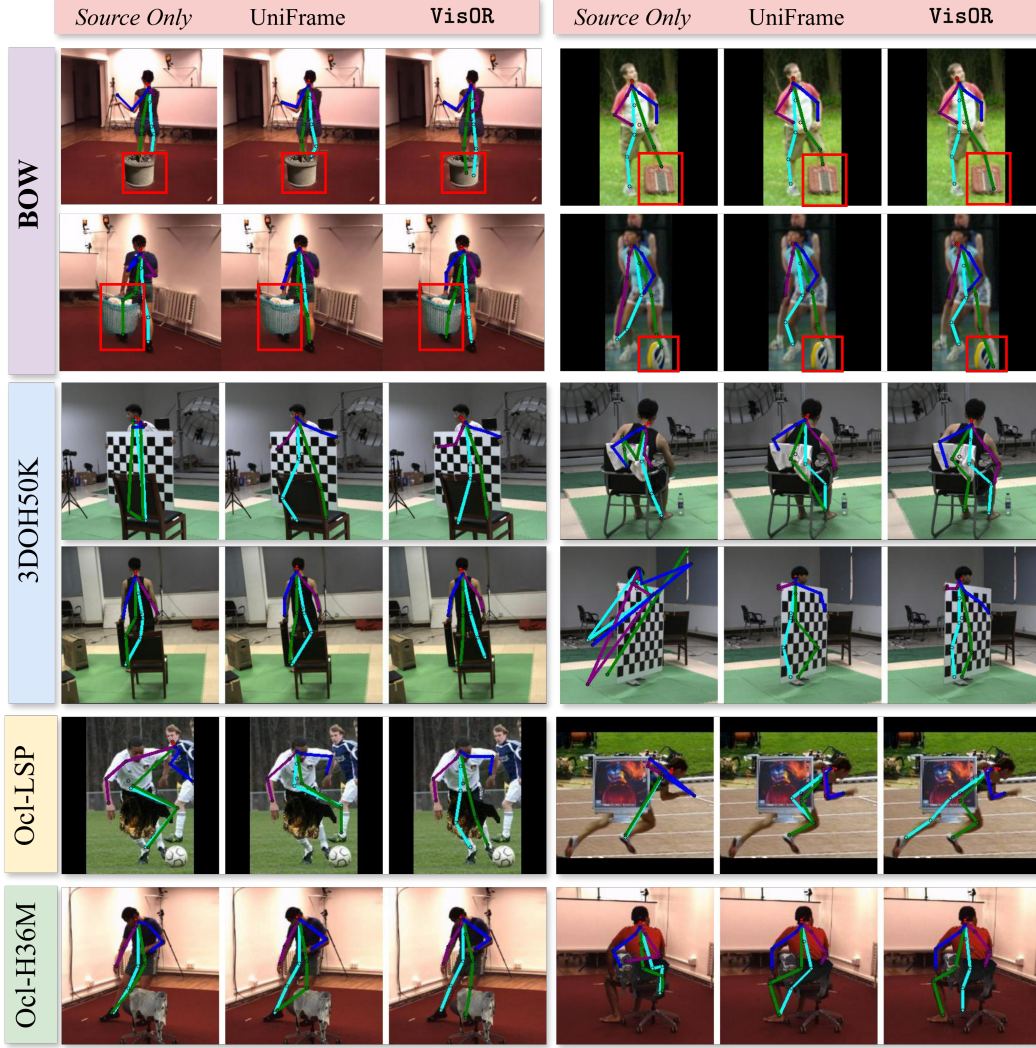


Figure 4. **Qualitative Results for Surreal→BOW, 3DOH50K, Ocl-LSP and Ocl-H36M.** Note that the best-performing baseline algorithm, UniFrame [1], performs poorly under occlusions, whereas **VisOR** generates accurate human poses.

$\mathcal{L}_{src}^{ocl}$	$\mathcal{L}_{pred}$	$\mathcal{L}_{ant}$	$\mathcal{L}_{vis}$	Avg. PCK@0.05
✗	✗	✗	✗	57.1
✓	✓	✗	✗	72.5
✓	✓	✓	✗	74.2
✓	✓	✓	✓	76.9

Table 5. Ablation study for **VisOR** in Surreal → 3DOH50K. method on the → 3DOH50K benchmark. We first observe that applying prediction-space consistency on the unlabeled target domain, combined with occlusion augmentations on the labeled source data (Eqn. 3 and Eqn. 4), enables **VisOR** to surpass the state-of-the-art UniFrame algorithm by approximately 3%. Adding regularization using a human pose prior (Eqn. 8) further boosts the performance to around 5%. Finally, introducing visibility-guided curriculum learning (Eqn. 10) leads to a total improvement of 7% over UniFrame, establishing new state-of-the-art results for unlabeled, occluded target domains.

6. Conclusion

We introduce **VisOR**, a self-supervised framework for robust human pose estimation under occlusions and domain shifts. Combining iterative pseudo-label refinement, a visibility-guided curriculum, and a learned pose prior, **VisOR** mitigates occlusion in unlabeled target data and outperforms existing methods on challenging occluded-pose benchmarks. We also propose **BOW**, a synthetic yet realistic benchmark for evaluating occlusion robustness in diverse indoor and outdoor environments. Across extensive experiments, **VisOR** consistently surpasses prior methods, setting a new state-of-the-art benchmark for self-supervised human pose estimation under occlusions.

Acknowledgments. This work was partially supported by NSF grants CMMI-2326309 and CNS-2312395.

References

- [1] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. In *ECCV*, 2022. 1, 2, 3, 4, 5, 7, 8
- [2] Janguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *CVPR*, 2021. 1, 2, 3, 4, 7
- [3] Qiang Zhou, Shiyin Wang, Yitong Wang, Zilong Huang, and Xinggang Wang. Human de-occlusion: Invisible perception and recovery for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3691–3701, 2021. 1
- [4] Rajat Modi, Vibhav Vineet, and Yogesh Rawat. On occlusions in video action detection: Benchmark datasets and training recipes. *Advances in Neural Information Processing Systems*, 36:57306–57335, 2023. 3
- [5] Rohit Lal, Saketh Bachu, Yash Garg, Arindam Dutta, Calvin-Khang Ta, Hannah Dela Cruz, Dripta S Raychaudhuri, M Salman Asif, and Amit K Roy-Chowdhury. Stride: Single-video based temporally continuous occlusion robust 3d pose estimation. *Winter Conference on Applications of Computer Vision*, 2025. 1
- [6] Tao Wang, Lei Jin, Zhang Wang, Xiaojin Fan, Yu Cheng, Yinglei Teng, Junliang Xing, and Jian Zhao. Decenternet: Bottom-up human pose estimation via decentralized pose representation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1798–1808, 2023. 2
- [7] Yiming Bao, Xu Zhao, and Dahong Qian. Fusepose: Imu-vision sensor fusion in kinematic space for parametric human pose estimation. *IEEE Transactions on Multimedia*, 2022.
- [8] Xipeng Chen, Junzheng Zhang, Keze Wang, Pengxu Wei, and Liang Lin. Multi-person 3d pose estimation with occlusion reasoning. *IEEE Transactions on Multimedia*, 26:878–889, 2024. 2
- [9] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 2, 3
- [10] Lumin Xu, Sheng Jin, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Zoomnas: searching for whole-body human pose estimation in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5296–5313, 2022.
- [11] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 2
- [12] 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, jul 2014. 2
- [13] 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. 2, 5
- [14] Alice Xiang. Being’seen’vs.’mis-seen’: Tensions between privacy and fairness in computer vision. *Harvard Journal of Law & Technology*, 36(1), 2022. 2
- [15] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. 2
- [16] Dripta S Raychaudhuri, Calvin-Khang Ta, Arindam Dutta, Rohit Lal, and Amit K Roy-Chowdhury. Prior-guided source-free domain adaptation for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14996–15006, 2023.
- [17] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020.
- [18] Adrian Lopez Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency. *arXiv preprint arXiv:1911.10033*, 2019.
- [19] Luan Tran, Kihyuk Sohn, Xiang Yu, Xiaoming Liu, and Manmohan Chandraker. Gotta adapt’em all: Joint pixel and feature-level domain adaptation for recognition in the wild. In *CVPR*, 2019.
- [20] Yuhe Ding, Jian Liang, Bo Jiang, Aihua Zheng, and Ran He. Maps: A noise-robust progressive learning approach for source-free domain adaptive keypoint detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [21] 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 4826–4836, 2023. 2
- [22] Qucheng Peng, Ce Zheng, Zhengming Ding, Pu Wang, and Chen Chen. Exploiting aggregation and segregation of representations for domain adaptive human pose estimation. *arXiv preprint arXiv:2412.20538*, 2024. 2, 3, 4, 7
- [23] Lu Zhou, Yingying Chen, Yunze Gao, Jinqiao Wang, and Hanqing Lu. Occlusion-aware siamese network for human pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 396–412. Springer, 2020. 2, 3
- [24] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10631–10638, 2020. 2, 3
- [25] Qihao Liu, Yi Zhang, Song Bai, and Alan Yuille. Explicit occlusion reasoning for multi-person 3d human pose estimation. In *European Conference on Computer Vision*, pages 497–517. Springer, 2022. 2, 3
- [26] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 2017. 2, 4
- [27] Yan Dai, Beitao Chen, Lianli Gao, Jingkuan Song, and Heng Tao Shen. DMH-CL: dynamic model hardness based curriculum learning for complex pose estimation. *IEEE Transactions on Multimedia*, 26:3180–3193, 2024. 2
- [28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 6

- [29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. [2](#), [6](#)
- [30] Arindam Dutta, Rohit Lal, Dripta S Raychaudhuri, Calvin-Khang Ta, and Amit K Roy-Chowdhury. Poise: Pose guided human silhouette extraction under occlusions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6153–6163, 2024. [2](#), [6](#)
- [31] Iason Sarandi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. How robust is 3d human pose estimation to occlusion? In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10302–10309. IEEE, 2020. [2](#), [3](#)
- [32] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Aberystwyth, UK, 2010. [2](#), [6](#)
- [33] 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, 2013. [2](#), [6](#)
- [34] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019. [3](#)
- [35] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. [3](#)
- [36] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.
- [37] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018.
- [38] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. [3](#)
- [39] 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. [3](#)
- [40] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *ECCV*, 2020.
- [41] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. *arXiv preprint arXiv:2104.02300*, 2021. [3](#)
- [42] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 488–504. Springer, 2020. [3](#)
- [43] Junjie Wang, Zhenbo Yu, Zhengyan Tong, Hang Wang, Jinxian Liu, Wenjun Zhang, and Xiaoyan Wu. Ocr-pose: Occlusion-aware contrastive representation for unsupervised 3d human pose estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5477–5485, 2022. [3](#)
- [44] Filipa Lino, Carlos Santiago, and Manuel Marques. 3d human pose estimation with occlusions: Introducing blendmimic3d dataset and gcnn refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4646–4656, June 2024. [3](#)
- [45] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, June 2021. [3](#)
- [46] Zhiyu Pan, Zhicheng Zhong, Wenxuan Guo, Yifan Chen, Jianjiang Feng, and Jie Zhou. Licampos: Combining multi-view lidar and rgb cameras for robust single-timestamp 3d human pose estimation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 2484–2494, February 2025. [3](#)
- [47] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9498–9507, 2019. [4](#)
- [48] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. [4](#)
- [49] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In *CVPR*, 2021. [5](#)
- [50] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *ECCV*, 2022. [5](#)
- [51] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. [5](#), [6](#), [7](#)
- [52] Riccardo Gatto and Sreenivasa Rao Jammalamadaka. The generalized von mises distribution. *Statistical Methodology*, 4(3):341–353, 2007. [5](#)
- [53] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020. [5](#)
- [54] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. [5](#)
- [55] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2020. [7](#)
- [56] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. [6](#)
- [57] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023. [6](#)
- [58] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. [7](#)