한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 32, No. 2, P. 33~45

Pose-Guided Reconstruction of Partially Visible Pedestrians under Vehicle Occlusion

Hyunjun Jang, Jae-Chern Yoo*

Dept. of Electrical and Computer Engineering, Sungkyunkwan University

10hour@skku.edu, *yoojc@skku.edu

차량 가림 환경에서 부분적으로 가시적인 보행자의 포즈 유도 재구성

장현준, 유재천*

성균관대학교 전자전기컴퓨터공학과

Abstract

This paper presents a pose-guided reconstruction framework that recovers full-body pose and silhouette from minimal lower-limb cues visible through the gap beneath an occluding vehicle. The implemented and evaluated pipeline consists of Occluded Lower-Limb Pose Detection (OLPD), Full-Body Pose Inference (FBPI), and Pose-Guided Silhouette Reconstruction (PGSR), which detect visible lower-limb keypoints, infer full-body keypoints, and generate a silhouette probability mask. We also introduce Minimum Required Visibility (MRV) to quantify the minimum visible body ratio required to satisfy a predefined reconstruction criterion. Under the adopted IoU $\geq$ 0.80 and Dice/F1 $\geq$ 0.90 criterion, the lowest visible ratio among successful test samples was 12.87%, and the measured latency of the FBPI + PGSR reconstruction path was 8.676 ms per sample, excluding OpenPose BODY_25 inference, vehicle-mask prediction, keypoint filtering, and HSV. OLPD results should be interpreted as a mask-given synthetic upper-bound because the current implementation assumes access to the synthetic vehicle-layer mask produced by the compositing pipeline. The reported 8.676 ms should be interpreted as a local, hardware-unspecified measurement of the FBPI+PGSR reconstruction path only. It excludes OpenPose BODY_25 inference, vehicle-mask prediction, keypoint filtering, and HSV, and therefore does not establish end-to-end real-time performance. Human Silhouette Verification (HSV) is not implemented or evaluated in the current paper and is discussed only as a prospective future extension.

요 약

본 논문은 차량에 의해 부분적으로 가려진 보행자를 대상으로, 차량 하부 틈새를 통해 관측되는 최소한의 하지 단서만으로 전신 자세와 실루엣을 복원하는 pose-guided reconstruction framework를 제안한다. 구현 및 평가된 파이프라인은 OLPD, FBPI, PGSR의 3단계로 구성되며, 가시적인 하지 keypoint를 검출한 뒤 전신 keypoint를 추론하고 최종적으로 silhouette probability mask를 생성한다. 또한 사전 정의된 복원 기준을 만족하는 데 필요한 최소 가시 비율을 정량화하기 위해 MRV를 도입한다. 통제된 synthetic paired dataset 실험에서, 채택한 IoU $\geq$ 0.80 및 Dice/F1 $\geq$ 0.90 기준하에 성공한 테스트 샘플들 중 가장 낮은 가시 비율은 12.87%였으며, OpenPose BODY_25 추론, vehicle-mask prediction, keypoint filtering, HSV를 제외한 측정된 FBPI+PGSR 복원 경로 기준 샘플당 8.676 ms의 처리 시간을 보였다. 현재 구현에서 OLPD는 compositing pipeline이 생성한 synthetic vehicle-layer mask에 접근한다고 가정하므로, OLPD 결과는 mask-given synthetic upper-bound로 해석되어야 한다. 보고된 8.676 ms는 하드웨어 정보가 별도로 기록되지 않은 로컬 환경에서 측정된 FBPI+PGSR 복원 경로 전용 수치로 해석되어야 하며, 전체 end-to-end 실시간 성능을 입증하지 않는다. HSV는 현재 구현 또는 평가에 포함되지 않았으며, 향후 확장 모듈로 별도 논의한다.

Keywords: Occluded Pedestrian Reconstruction, Human Pose Estimation, Full-Body Pose Inference, Pose-Guided Silhouette Reconstruction, Traffic Safety Perception.

키워드: 차량 가림 보행자 재구성, 인체 자세 추정, 전신 자세 추론, 포즈 유도 실루엣 재구성, 교통 안전 인식

*corresponding author: Jae-Chern Yoo / Dept. of Electrical and Computer Engineering, Sungkyunkwan University (yoojc@skku.edu)

Received : 2026.04.01./ Review completed : 1st 2026.04.16. / Accepted : 2026.04.30.
DOI : 10.15701/kcgs.2026.32.2.33
ISSN : 1975-7883(Print)/2383-529X(Online)

1. Introduction

Roadside pedestrian accidents often occur when a pedestrian suddenly emerges from behind a parked vehicle before sufficient visual evidence becomes available to onboard perception modules. This risk is particularly high in dense roadside-parking environments, where only a narrow under-vehicle gap may reveal limited lower-limb cues.

Existing pose estimators and occlusion-recovery methods do not directly address full-body silhouette reconstruction from extremely sparse under-vehicle observations [1, 2]. In parallel, ghost-probe approaches primarily focus on hazard-zone prediction rather than explicit shape-level recovery from partial visual evidence. This leaves a practical gap between visible-cue detection and reconstruction-oriented reasoning for severely occluded pedestrians.

We propose a pose-guided reconstruction framework whose implemented and evaluated pipeline comprises OLPD, FBPI, and PGSR for recovering full-body structure and silhouette from sparse lower-limb cues beneath parked vehicles. A separate HSV stage is treated only as a prospective future extension and is not implemented or evaluated in this paper.

The main contributions are as follows. First, we formulate partial under-vehicle pedestrian visibility as a pose-guided reconstruction problem. Second, we introduce an implemented three-stage cascade for recovering full-body pose and silhouette from sparse lower-limb cues. Third, we introduce MRV to quantify occlusion tolerance. Fourth, we construct a synthetic paired dataset for controlled evaluation.

2. Related Work

2.1 Occlusion-Aware Human Pose and Mesh Recovery

OpenPose (Cao et al., 2017) and CrowdPose (Li et al., 2019) provide the visible-keypoint estimation foundation of our framework [1, 2]. More recent work has pushed pose and mesh reasoning under occlusion by explicitly modeling part visibility, spatial context, diffusion priors, or richer whole-body structure. For example, PARE uses body-part-guided attention to improve 3D human body estimation under partial occlusion [3]; visibility-aware transformer reasoning improves 2D pose estimation under occlusion by suppressing unreliable occluder features [4]; Occluded Human Mesh Recovery (OCHMR) augments top-down mesh recovery with spatial context for person-person occlusion [5]; AiOS moves toward an all-in-one-stage expressive pose-and-shape pipeline without an external detector [6]; DPMesh exploits diffusion priors for occluded human mesh recovery [7]; and ScoreHMR uses score-guided diffusion for 3D human recovery across several inverse-problem settings [8]. These studies demonstrate strong recovery capability when broader human extent and richer contextual evidence are available. However, they are not direct one-to-one solutions for the present setting, because our target problem is under-vehicle occlusion with only minimal lower-limb cues and a lightweight staged reconstruction output of full-body keypoints and silhouette probability masks rather than full 3D meshes.

In the KCGS literature, K-SMPL, silhouette-driven human-shape estimation, sensor-based motion reconstruction, and low-cost interactive motion-capture systems provide useful representation- and motion-level context [9, 10, 11, 12], but they are not designed for under-vehicle occlusion or minimal lower-limb cue recovery.

2.2 Human De-Occlusion and Amodal Completion

Human de-occlusion and amodal completion address the inference of invisible structure behind occluders. Earlier human-specific work includes object-occluded human shape and pose recovery [13], human de-occlusion with invisible-mask and appearance recovery [14], and 2D pose-guided complete silhouette estimation for occluded human bodies [15]. Subsequent work has expanded both realism and scope: OccNeRF studies human rendering from object-occluded monocular videos using geometry and visibility priors [16]; Amodal Completion via Progressive Mixed Context Diffusion uses context-guided diffusion for completion [17]; pix2gestalt synthesizes wholes for zero-shot amodal segmentation [18]; and Amodal Ground Truth and Completion in the Wild broadens amodal benchmarking and completion in real images [19]. Complementary visible-mask extraction backbones such as Mask2Former and SAM strengthen general-purpose segmentation quality [20, 21]. These methods are highly relevant because they infer invisible masks, shapes, or appearance from partial observations or strengthen the visible mask extraction on which downstream completion may depend. Nevertheless, most of them target generic amodal or appearance completion, broader segmentation, or settings with richer visible context than ours, and are not explicitly formulated as lightweight staged reconstruction from minimal lower-limb cues under under-vehicle occlusion.

2.3 Pose Completion and Lightweight Staged Reconstruction

A closer family to FBPI is pose completion from partial observations. SDR-GAIN targets high-real-time occluded pedestrian pose completion for autonomous driving [22]; LInKs adopts a lift-then-fill strategy for partial-pose lifting under occlusion [23]; Cross-view and Cross-pose Completion studies missing-pose recovery for 3D human understanding [24]; and HiPART performs hierarchical densification for occluded 3D pose estimation while improving robustness and computational efficiency relative to heavier generative alternatives [25]. These works are important because they show that missing skeletal structure can be inferred from incomplete pose evidence. However, prior pose-completion methods generally target 3D lifting or broader partial-body settings rather than the present combination of under-vehicle occlusion, minimal lower-limb cues, and lightweight staged reconstruction. By contrast, our pipeline explicitly decomposes the task into OLPD, FBPI, and PGSR so that visible-cue extraction, keypoint inference, and silhouette reconstruction remain modular. This decomposition is not presented as outperforming prior completion families; rather, it targets a narrower deployment-motivated regime in which only limited under-vehicle evidence is available. Consistent with the rest of the paper, OLPD results are interpreted only as mask-given synthetic upper-bound filtering under a known occlusion mask, FBPI metrics are interpreted as consistency with paired OpenPose reference poses rather than manual human pose accuracy, and the current validation remains synthetic and child-pedestrian-limited.

2.4 Traffic-Safety Occlusion Reasoning

In traffic-safety research, occlusion is often addressed as a ghost-probe or hazard-prediction problem rather than as explicit human reconstruction. *DPGP: A Hybrid 2D-3D Dual Path Potential Ghost Probe Zone Prediction Framework for Safe Autonomous Driving* (Qu et al., 2025), for example, addresses ghost-probe zone prediction while emphasizing the practical cost of specialized sensors and computation [26]. This line of work is highly relevant because it treats occlusion as a safety-critical perception problem. However, the target output differs from ours: ghost-probe methods typically predict risk zones, motion intent, or warning cues, whereas the present study focuses on reconstructing full-body keypoints and a silhouette probability mask from minimal lower-limb cues under under-vehicle occlusion. Our method should therefore be viewed as complementary to traffic-safety occlusion reasoning rather than as a replacement for end-to-end warning systems.

3. Method

To address the target problem, we introduce a pose-guided reconstruction framework that separates the implemented pipeline from a prospective future extension. The implemented and evaluated pipeline consists of OLPD, FBPI, and PGSR. A separate HSV stage is discussed in Section 3.4 only as a future extension and is not part of the implemented or evaluated system in this paper.

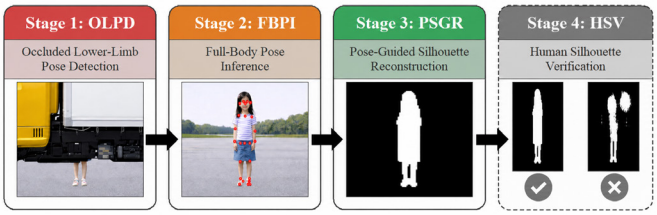

Figure 1: Overview of the framework. Implemented pipeline: OLPD, FBPI, and PGSR. Future extension / prospective HSV: a conceptual verification stage that was not implemented or evaluated in this paper.

In Fig. 1, only the OLPD–FBPI–PGSR path belongs to the implemented pipeline, whereas HSV is shown separately as a prospective future extension.

3.1 Occluded Lower-Limb Pose Detection (OLPD)

In the first stage, termed Occluded Lower-Limb Pose Detection (OLPD), the goal is to extract only observable lower-limb cues from vehicle-occluded pedestrian images, where most of the body is hidden by a vehicle and only a small portion of the legs is visible through the under-vehicle gap. To this end, we apply a pre-trained OpenPose BODY_25 model to the full input image without additional fine-tuning.

In the current synthetic dataset, the vehicle layer mask is directly available from the compositing process and is used as the occlusion mask in OLPD. Joints that fall inside this vehicle-occluded region are removed before retaining the visible lower-limb subset. The estimated 2D pose is then passed to partial lower-limb pose extraction, which retains only the visible lower-limb joints defined in BODY_25. The final output is a partial pose representation expressed as visible lower-limb keypoints. Because upper-body evidence is mostly unavailable in this setting, OLPD acts as a preprocessing stage that stabilizes downstream inference by isolating these limited yet reliable lower-limb observations for the next stage. In the current implementation, OLPD assumes access to the synthetic vehicle-layer mask produced by the compositing pipeline. Therefore, OLPD results do not measure automatic vehicle segmentation or deployment-time perception. Accordingly,

any reported OLPD success rate should be interpreted as a mask-given synthetic upper-bound for lower-limb keypoint filtering under a known occlusion mask. Any OLPD-related latency reported later refers only to keypoint filtering after OpenPose outputs are available; it does not measure OpenPose BODY_25 inference itself. Section 4.2 separately discusses the oracle-mask setting that uses the synthetic vehicle-layer mask and an illustrative predicted-mask setting that substitutes an externally estimated mask for qualitative comparison only.

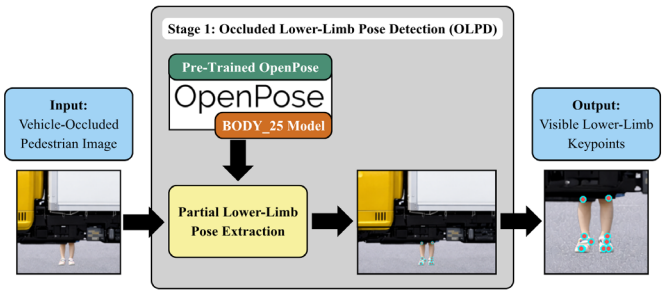

Figure 2: Stage 1: Occluded Lower-Limb Pose Detection (OLPD) in the mask-given synthetic setting. A vehicle-occluded pedestrian image is processed by a pre-trained OpenPose BODY_25 model for partial lower-limb pose extraction using the oracle synthetic vehicle mask available from the compositing pipeline, producing visible lower-limb keypoints. This figure illustrates a mask-given synthetic upper-bound filtering setup and does not evaluate automatic vehicle segmentation or deployment-time perception.

3.2 Full-Body Pose Inference (FBPI)

In the second stage, termed Full-Body Pose Inference (FBPI), the objective is to reconstruct a complete human pose from the sparse lower-limb observations extracted in Stage 1. As illustrated in Fig. 3, the module takes visible lower-limb keypoints as input and predicts completed full-body keypoints through PartialToFullKeypointNet. To train this module, paired synthetic samples were constructed using two image versions of the same scene: a non-occluded reference image and a synthetic occlusion image generated by car-mask-based composition. Specifically, the dataset was created using Gemini-based image generation followed by manual three-layer composition of background, pedestrian, and vehicle elements. A total of 1,000 paired samples were prepared and split into 800 training, 100 validation, and 100 test samples at the paired-sample-ID level. The occluded image and the non-occluded reference image from the same pair were always assigned to the same split. In the current dataset, paired sample IDs were ordered by index and then partitioned into the 800/100/100 split without separating the two images of a pair. For each pair, OpenPose was applied to both the non-occluded image and the synthetically occluded image. The pose detected from the synthetically occluded image was used as the partial input, whereas the pose detected from the corresponding non-occluded image was used as the paired OpenPose reference pose for full-body keypoint inference. The target full-body keypoints are not manually annotated human pose ground truth. They are OpenPose BODY_25 outputs extracted from the paired non-occluded reference image. Therefore, FBPI learns to predict keypoints consistent with the paired OpenPose reference pose rather than with manual keypoint annotation.

Let the partial pose estimated from the occluded image be denoted by

$$\mathbf{K} \in \mathbb{R}^{25\times 3}, \quad \mathbf{K}_j = [x_j, y_j, c_j], \tag{1}$$

where x_j, y_j, and c_j denote the normalized 2D coordinates and confidence of the j-th BODY_25 joint. The coordinates are normalized with respect to the image width W and height H as

$$x_j^{\text{norm}} = \frac{x_j^{\text{pixel}}}{W}, \qquad y_j^{\text{norm}} = \frac{y_j^{\text{pixel}}}{H}. \tag{2}$$

In addition, a binary visibility vector

$$\mathbf{v} \in \{0,1\}^{25}, \quad v_j \in \{0,1\} \tag{3}$$

is used to indicate whether each joint remains observable after occlusion masking. For occluded joints, the corresponding entries are set to $x_j = 0$, $y_j = 0$, $c_j = 0$, and $v_j = 0$. Therefore, the model always receives a fixed-size pose structure with 25 joint slots, while the effective amount of valid information varies depending on the number of visible lower-limb joints in each sample.

For each joint, the input token is constructed by concatenating coordinate, confidence, and visibility information:

$$\mathbf{z}_j = [x_j, y_j, c_j, v_j] \in \mathbb{R}^4, \quad \mathbf{Z} \in \mathbb{R}^{25\times 4}. \tag{4}$$

These 25 input joint tokens are then projected into a latent space and processed by a transformer encoder with learnable joint positional embeddings. As shown in Fig. 4, the FBPI network, referred to as PartialToFullKeypointNet, is implemented as a transformer-based full-body inference network with hidden dimension 256, 4 transformer encoder layers, 8 attention heads, feedforward dimension 1024, GELU activation, and dropout rate 0.1 [27]. The network outputs a completed full-body pose

$$\hat{\mathbf{Y}} \in \mathbb{R}^{25\times 3}, \quad \hat{\mathbf{Y}}_j = [\hat{x}_j, \hat{y}_j, \hat{c}_j]. \tag{5}$$

Although the output tensor has shape 25×3, the training objective focuses on coordinate regression against the OpenPose-derived

reference pose for the completed pose. Let

$$\hat{\mathbf{p}}_j = [\hat{x}_j, \hat{y}_j], \quad \mathbf{p}_j^{\text{ref}} = [x_j^{\text{ref}}, y_j^{\text{ref}}], \tag{6}$$

where $\mathbf{p}_j^{\text{ref}}$ is taken from the paired OpenPose reference pose extracted from the non-occluded image. The default training loss is then a confidence-weighted L1 loss with respect to this OpenPose-derived reference pose:

$$\mathcal{L}_{xy} = \frac{\sum_{j=1}^{25} w_j \left\| \hat{\mathbf{p}}_j - \mathbf{p}_j^{\text{ref}} \right\|_1}{\sum_{j=1}^{25} w_j + \epsilon}, \qquad w_j = c_j^{\text{ref}}. \tag{7}$$

Thus, higher-confidence reference joints contribute more to the loss. Confidence and visibility are not supervised separately; visibility is used only to mark observed versus missing joints at the input. Accordingly, FBPI should be interpreted as consistency with the paired OpenPose reference pose rather than as direct human pose accuracy.

To preserve reliable observations while only hallucinating missing parts, the final completed pose $\mathbf{Y}$ is obtained by residual fusion between the original partial pose and the predicted pose:

$$\mathbf{Y}_j = v_j \mathbf{K}_j + (1 - v_j)\hat{\mathbf{Y}}_j. \tag{8}$$

This formulation preserves visible joints and predicts only occluded ones. After completion, reconstructed occluded joints are assigned confidence 1 before the pose is passed to the next stage. The current experiments are limited to single-person samples; multi-person extension remains future work. A transformer encoder is used to model long-range dependencies between visible lower-limb joints and missing upper-body joints within a fixed skeletal topology.

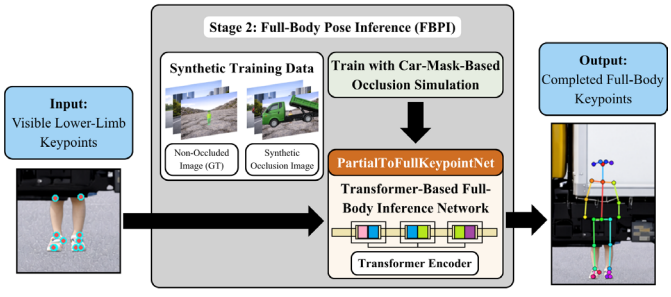

Figure 3: Stage 2: Full-Body Pose Inference (FBPI). Input: visible lower-limb keypoints. Synthetic training data are formed by paired non-occluded reference image and synthetic occlusion image samples, and the model is trained with car-mask-based occlusion simulation to output completed full-body keypoints. Here, "GT" in the schematic denotes the paired OpenPose reference pose extracted from the non-occluded image, not manually annotated human pose ground truth.

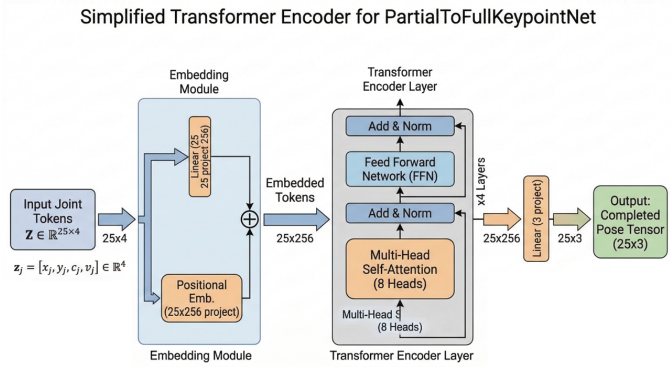

Figure 4: Simplified transformer encoder for PartialToFullKeypointNet. Input joint tokens $\mathbf{Z} \in \mathbb{R}^{25\times4}$ are embedded and processed by 4 transformer encoder layers with multi-head self-attention (8 heads), followed by linear projection to a completed pose tensor (25×3).

3.3 Pose-Guided Silhouette Reconstruction (PGSR)

In the third stage, Pose-Guided Silhouette Reconstruction (PGSR) recovers a full human silhouette mask from the completed full-body pose without RGB appearance cues. The implemented KeypointToMaskNet is a U-Net-based keypoint-to-mask network [28] trained on paired OpenPose reference poses from the non-occluded image and the corresponding manually isolated silhouette masks, while inference uses only the completed pose predicted by FBPI. Because training uses cleaner reference poses whereas inference uses FBPI outputs, PGSR should be interpreted as a pose-conditioned upper-bound with respect to pose-input quality.

Let the completed full-body pose passed from FBPI at inference time be denoted by

$$\mathbf{Y} \in \mathbb{R}^{25\times3}, \quad \mathbf{Y}_j = [x_j, y_j, c_j], \tag{9}$$

where x_j, y_j, and c_j are the normalized 2D coordinates and confidence of the j-th BODY_25 joint. The stacked pose matrix is written as

$$\mathbf{Y} = \left[\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, \ldots, \mathbf{Y}_{25}^\top\right]^\top. \tag{10}$$

During PGSR training, however, the implemented model uses the paired OpenPose reference full-body pose from the non-occluded image, denoted by $\mathbf{Y}^{\text{ref}} \in \mathbb{R}^{25\times3}$, rather than the residual-fused FBPI output $\mathbf{Y}$. At the level preserved by the surviving experiment record, the pose-conditioned PGSR mapping can be written as

$$\hat{\mathbf{M}} = \mathcal{G}(\mathbf{P}), \quad \hat{\mathbf{M}} \in [0,1]^{256\times256}. \tag{11}$$

Here, $\mathbf{P} = \mathbf{Y}^{\text{ref}}$ during training and $\mathbf{P} = \mathbf{Y}$ during inference. The network follows a U-Net-style encoder-decoder topology with

skip connections, base channel width 64, and a final sigmoid output at 256×256 resolution. The silhouette mask is a supervision target only; it is not part of the inference input. However, the surviving manuscript assets do not preserve the exact internal tensor-to-spatial projection that converts the 25×3 pose representation into the first spatial feature map consumed by the encoder. Therefore, the current paper documents PGSR at the level of pose conditioning, U-Net-style topology, and output resolution, but not as a fully layer-by-layer reproducible architecture specification. Because $\mathbf{Y}^{\mathrm{ref}}$ and $\mathbf{Y}$ are not drawn from exactly the same pose distribution, the reported PGSR numbers should not be read as the performance of a model trained on the exact deployment-time pose input.

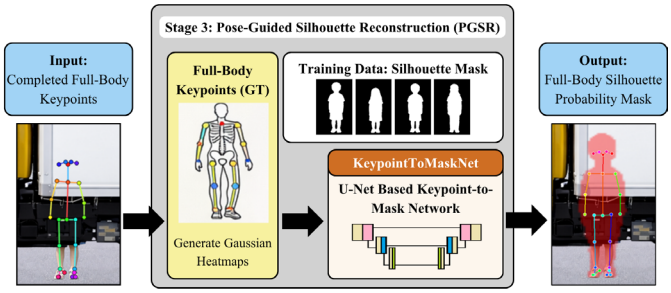

Figure 5: Stage 3: Pose-Guided Silhouette Reconstruction (PGSR). The figure should be read by separating training and inference roles. During training, the input pose is the paired OpenPose reference full-body pose from the non-occluded sample, and the manually isolated binary silhouette mask is used as the supervision target for KeypointToMaskNet. During inference, the residual-fused completed pose predicted by FBPI is the only network input, and the output is a full-body silhouette probability mask. Thus, the current implementation contains a pose-input train–test mismatch: PGSR is trained on reference poses but evaluated downstream with FBPI outputs. The illustrated pose-conditioning block is schematic only. In particular, the current manuscript does not preserve the exact internal projection from the 25×3 pose representation to the first spatial feature map of the encoder, so this figure should not be interpreted as a layer-by-layer reproducible architecture diagram.

The supervision target is a cleaned, manually isolated full-body silhouette mask. Although the original mask is binary, resizing to the output resolution introduces soft boundary values that provide smoother edge supervision. PGSR is optimized with a compound loss combining focal binary cross-entropy and Dice loss:

$$\mathcal{L}_{\mathrm{PGSR}} = \mathcal{L}_{\mathrm{focal}} + \mathcal{L}_{\mathrm{dice}}. \tag{12}$$

This loss encourages both pixel-wise discrimination and global shape consistency. During training, it is applied directly to the probability map without hard thresholding; for evaluation, the predicted mask is binarized at 0.5.

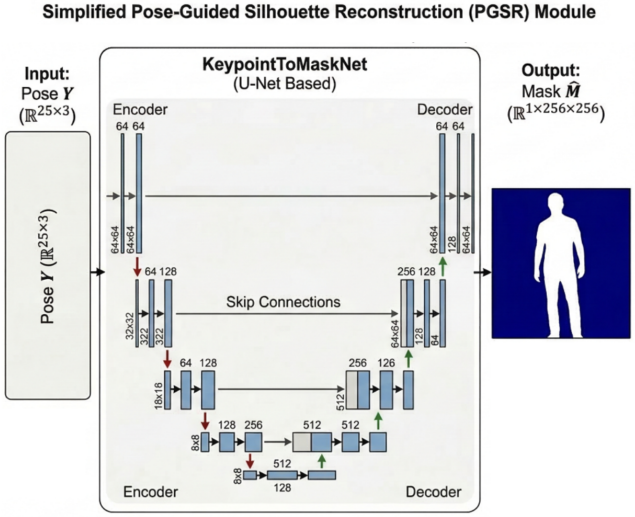

Figure 6: Conceptual schematic of the Pose-Guided Silhouette Reconstruction (PGSR) module. KeypointToMaskNet is documented here only as a U-Net-style encoder-decoder with skip connections that maps a pose-conditioned representation to a silhouette probability map $\hat{\mathbf{M}} \in \mathbb{R}^{1\times256\times256}$. The conditioning source is the paired OpenPose reference pose $\mathbf{Y}^{\mathrm{ref}}$ during training and the residual-fused FBPI output $\mathbf{Y}$ during inference. The exact internal projection from the 25×3 pose tensor to the first spatial encoder feature map is not preserved in the current manuscript assets, so this figure is schematic rather than a fully reproducible layer specification.

The PGSR module is trained on the same 1,000 paired synthetic samples used for Stage 2, with the same paired-sample-level 800/100/100 split. That is, the occluded image, the paired non-occluded reference image, and all derived supervision from the same pair always belong to the same partition. Each sample contains a single pedestrian instance, so the model learns a one-person mapping from full-body pose to silhouette mask. The present experiments are limited to single-person samples. Extension to multi-person scenes would require reliable person-wise separation before pose completion and is left for future study. By reconstructing a dense silhouette from the inferred skeletal configuration, PGSR provides a shape-level representation that could be used by a future verification module.

3.4 Prospective Human Silhouette Verification (HSV)

As a prospective extension to the implemented three-stage pipeline, Human Silhouette Verification (HSV) would take the PGSR output mask and reject morphologically implausible reconstructions before downstream risk-assessment or warning modules. This module was not implemented or quantitatively evaluated in the present study.

One possible realization is to transfer silhouette-validity labels from a multimodal teacher model to a lightweight mask-based student classifier for efficient deployment. Such a verifier could help suppress fragmented masks, merged foreground regions, and non-human topology. However, because no HSV-specific dataset, teacher-labeling protocol, or evaluation benchmark was established, we report no HSV metrics in this paper.

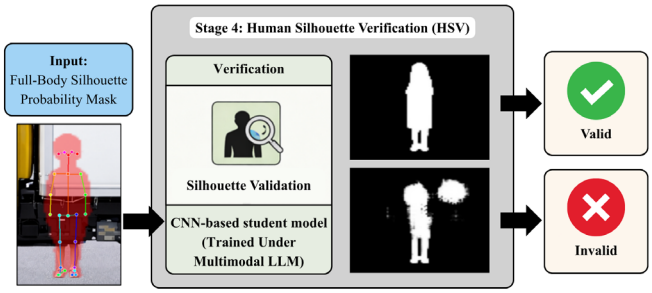

Figure 7: Conceptual illustration of a prospective Human Silhouette Verification (HSV) module. This module is not implemented or quantitatively evaluated in the present study.

4. Experimental Results

4.1 Dataset

To construct a dataset suitable for reconstruction experiments on pedestrians partially visible beneath parked vehicles, we built a fully synthetic image dataset rather than relying on real-world data. Collecting real images for this scenario is practically difficult because the target condition requires a pedestrian to be visible only through the narrow gap between a vehicle and the ground. Therefore, we generated source visual assets using Gemini-based image generation and manually composited them into final training samples.

As summarized in Fig. 8, the dataset was built from three asset groups: *Background Image Dataset [10pcs]*, *Pedestrian (Kid) Image Dataset [10pcs]*, and *Vehicle Image Dataset [10pcs]*. Each pedestrian image was manually processed in Adobe Photoshop to remove the background and obtain a transparent foreground layer. The foreground pedestrian layer was then combined with a vehicle layer and a background layer using a classical layer-based compositing pipeline. From each scene configuration, we created two corresponding images: a non-occluded version and an occluded version in which the pedestrian was partially hidden by the vehicle. As a result, the dataset consisted of *Occluded Sample Image Dataset [1,000 paired - manually composited]*, yielding 2,000 images in total.

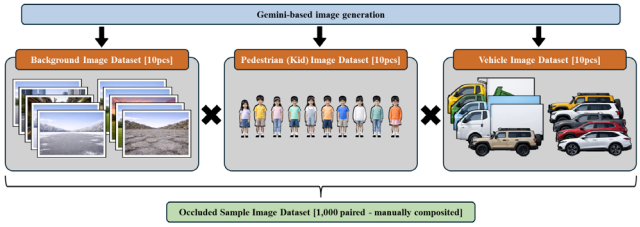

Figure 8: Gemini-based image generation pipeline for constructing the synthetic occluded dataset. The final *Occluded Sample Image Dataset [1,000 paired - manually composited]* is built by combining *Background Image Dataset [10pcs]*, *Pedestrian (Kid) Image Dataset [10pcs]*, and *Vehicle Image Dataset [10pcs]*.

In this study, an occluded sample is defined as a case in which only the lower body, mainly the legs visible through the gap beneath the vehicle, remains observable. Cases in which the pedestrian was completely invisible were excluded, because OpenPose could not reliably extract meaningful lower-body keypoints in such situations. The generated image resolution was 1536×1024 pixels.

For annotation, two types of supervision were prepared. First, a person silhouette mask was obtained from the manually isolated pedestrian foreground and used as the ground-truth mask for silhouette reconstruction. Second, full-body keypoints were extracted from the non-occluded images using the OpenPose BODY_25 model and used as the paired OpenPose reference pose for FBPI. During training, lower-body keypoints extracted from the occluded images served as the input representation, while the corresponding full-body keypoints from the paired non-occluded images were used as the reference output. No manual keypoint annotation was used for FBPI training or evaluation in the current paper. Because these pose targets were automatically generated from the paired non-occluded images rather than manually annotated, the reported FBPI performance should be interpreted as consistency with the paired OpenPose reference poses. Each pose sample was represented as a 25×3 matrix, where each of the 25 BODY_25 joints was described by its x-coordinate, y-coordinate, and confidence score.

The dataset was divided into 800 training pairs, 100 validation pairs, and 100 test pairs. The split was performed at the paired-sample-ID level after ordering the paired sample IDs by index. The occluded image and the non-occluded image belonging to the same

pair were always assigned to the same split; they were never separated across training, validation, and test partitions. Because the current synthetic construction used child-pedestrian foreground assets, the reported results should be interpreted as a controlled feasibility study under a child-pedestrian setting rather than as a universal benchmark for all pedestrian appearances.

4.2 Quantitative Results and Visual Comparison

Because HSV was not implemented, the quantitative results reported in this section correspond only to the implemented OLPD, FBPI, and PGSR components. Table 2 separates the measured latency scope. Because the original experiment log did not preserve the CPU/GPU model, inference runtime framework, batch size, or warm-up/repetition counts, a fully specified timing protocol cannot be reported retrospectively. The reported 8.676 ms should therefore be interpreted as a local, hardware-unspecified measurement of the FBPI + PGSR reconstruction path only. It excludes OpenPose BODY_25 inference, vehicle-mask prediction, keypoint filtering, and HSV. The 0.077 ms OLPD-related timing reported here does not measure OpenPose inference; it measures only keypoint filtering time after OpenPose outputs are available, using the known synthetic vehicle mask. When this post-OpenPose keypoint-filtering step is added to FBPI + PGSR, the combined measured latency becomes 8.753 ms, but this value still excludes OpenPose BODY_25 inference, vehicle-mask prediction, and HSV. No manual keypoint annotation was used for FBPI evaluation in this section. Accordingly, FBPI metrics should be interpreted as consistency with OpenPose-derived reference poses rather than as actual human pose accuracy. Moreover, because the dataset is constructed from only 10 background assets, 10 vehicle assets, and 10 child pedestrian assets, the reported numbers should be interpreted as a controlled feasibility study under synthetic conditions rather than as evidence of broad real-world recovery. In this section, oracle-mask OLPD results should be interpreted strictly as mask-given synthetic upper-bound results obtained with the synthetic vehicle-layer mask from the compositing pipeline; they do not measure automatic vehicle segmentation or deployment-time perception.

4.2.1 Oracle-Mask vs Predicted-Mask Comparison

Before comparing against external generic completion baselines, we first separate the oracle-mask setting from the predicted-mask setting. Fig. 9 compares these two regimes. In the oracle-mask setting, the synthetic vehicle-layer mask is provided directly by the compositing pipeline, so the resulting OLPD behavior should be interpreted only as a mask-given synthetic upper-bound rather than as automatic vehicle segmentation or deployment-time perception. By contrast, the predicted-mask panel in Fig. 9 is included only as an illustrative external segmentation example in which an externally estimated vehicle mask is substituted for the oracle mask and then passed to the same downstream reconstruction path. The exact external mask-prediction procedure used to prepare that illustrative panel is not documented in sufficient detail in the current manuscript assets and is not part of the implemented benchmarked pipeline reported in this paper. Accordingly, the predicted-mask setting in Fig. 9 should not be read as a reproducible model specification or formal method comparison. Moreover, the current manuscript does not provide quantitative evaluation for predicted vehicle-mask IoU, OLPD lower-limb F1 under the predicted mask, FBPI OKS under the predicted mask, PGSR IoU/Dice under the predicted mask, or separated mask-prediction latency. Therefore, the predicted-mask results in Fig. 9 should be interpreted as qualitative evidence only rather than as quantitative support for practical deployment.

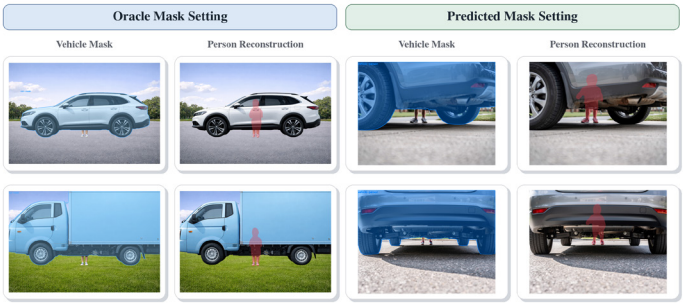

Figure 9: Qualitative comparison between the oracle-mask and predicted-mask settings. Left: the synthetic vehicle mask is provided directly by the compositing process and used as an oracle occlusion mask; this oracle-mask condition should be interpreted as a mask-given synthetic upper-bound. Right: an externally estimated vehicle mask is used in place of the oracle mask and the same downstream reconstruction path is applied. The predicted-mask panel is presented only as an illustrative external segmentation example; the exact mask-prediction procedure is not specified as part of the implemented benchmarked pipeline in the current manuscript. Because no predicted-mask quantitative evaluation is reported, this panel should be interpreted as qualitative evidence only.

4.2.2 Stage-Wise Sensitivity to Mask Errors

To analyze how vehicle-mask errors affect the OLPD–FBPI–PGSR pipeline, we applied four perturbation types: undersegmentation, oversegmentation, boundary jitter, and partial missing. Fig. 10 reports vehicle-mask IoU, OLPD lower-limb F1, FBPI OKS with respect to OpenPose-derived reference poses on occluded samples, and PGSR mask IoU across mild-to-extreme severity levels.

Undersegmentation, oversegmentation, and boundary jitter caused only limited downstream change even at extreme severity. By contrast, partial missing was much more harmful because it removed visible lower-limb evidence before OLPD and then propagated errors to FBPI and PGSR. Under extreme partial missing, vehicle-mask IoU remained 0.973, but OLPD lower-limb F1 fell to 0.854 and both FBPI OKS and PGSR IoU deviated strongly from the oracle baseline.

These results show that vehicle-mask IoU alone is not sufficient to explain downstream stability. Compared with contour drift or mild over-/under-segmentation, errors that erase visible lower-limb cues are more consequential because they directly remove the evidence used by OLPD. Accordingly, preserving lower-limb visibility evidence is more important in this pipeline than overall vehicle-mask overlap. The OLPD curves in Fig. 10 should therefore be interpreted as mask-given synthetic upper-bound sensitivity trends under a known occlusion mask rather than as automatic vehicle-segmentation performance.

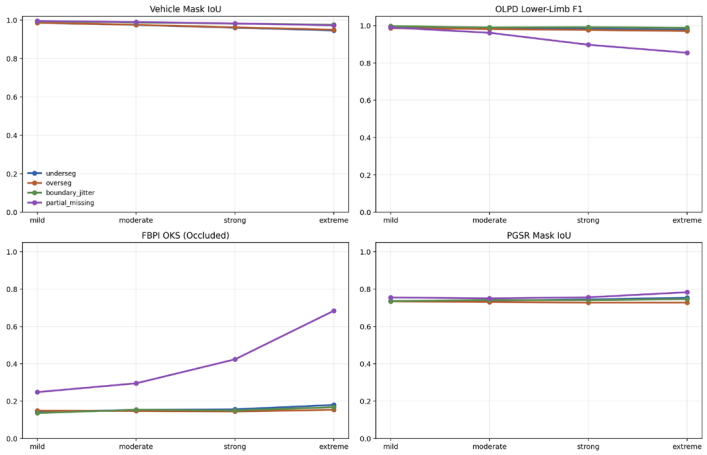

Figure 10: Stage-wise diagnostic curves under synthetic occlusion-mask perturbations around the oracle synthetic vehicle-mask condition. From top-left to bottom-right: vehicle-mask IoU, OLPD lower-limb F1, FBPI OKS with respect to OpenPose-derived reference poses on occluded samples, and PGSR mask IoU. Each curve corresponds to one perturbation family evaluated from mild to extreme severity. OLPD-related trends should be interpreted as mask-given synthetic upper-bound sensitivity analysis rather than automatic vehicle-segmentation performance, and FBPI should be interpreted as consistency with paired OpenPose reference poses rather than manual human pose accuracy.

4.2.3 Comparison with Generic Completion Baselines

We next strengthen the comparison beyond simple OpenPose-only or SAM-only references by additionally evaluating three generic full-body completion baselines: Stable Diffusion Inpainting, Stable Diffusion v1.5 img2img masked completion, and Simple LaMa inpainting. These methods operate on masked RGB completion rather than sparse lower-limb keypoints, and their completed RGB outputs are converted into person-mask hypotheses for comparison with our reconstructed silhouette. Therefore, the resulting comparison should be interpreted as a stronger cross-paradigm small-scale qualitative screening rather than as a perfectly stage-matched ablation.

Fig. 11 presents a qualitative comparison between OURS and these stronger generic full-body completion baselines on under-vehicle child samples.

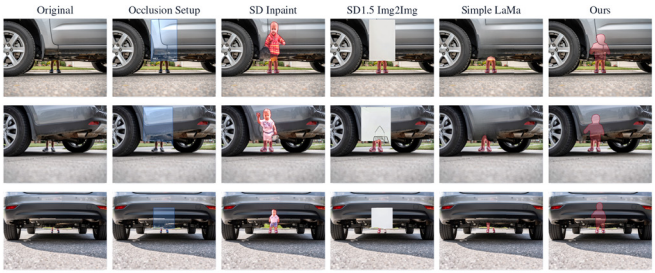

Figure 11: Qualitative comparison against stronger generic full-body completion baselines on under-vehicle child samples. Generic completion outputs are converted into person-mask hypotheses for comparison with OURS.

Table 1 summarizes a small-scale qualitative screening on three representative synthetic/composited under-vehicle child samples, focusing on illustrative upward completion behavior and runtime relevance. Here, the top lift ratio corresponds to the previously reported height ratio, so a larger value indicates a stronger upward lift of the completed body into the hidden region for these three examples only; it should not be interpreted as a quantitative benchmark.

Table 1: Executed generic full-body completion baselines, summarized for a small-scale qualitative screening by top lift ratio and runtime relevance.

Model	Top lift ratio	Runtime
OURS	3.835	8.676 ms (FBPI + PGSR only)
SD Inpaint	3.975	1.4125 s
SD1.5 Img2Img	0.977	1.1771 s
Simple LaMa	1.484	1.1824 s

Within this small-scale qualitative screening, SD Inpaint showed the largest top lift ratio on the three representative synthetic/composited examples, whereas OURS remained close and was the only method with a measured lightweight reconstruction-path latency (8.676 ms for FBPI + PGSR only, excluding OpenPose BODY_25 inference, vehicle-mask prediction, keypoint filtering, and HSV). Simple LaMa stayed closer to the visible evidence, and SD1.5 Img2Img was unstable. Overall, generic diffusion com-

pletion can produce strong upward lift, but the proposed staged pipeline remains more lightweight and structurally controlled for the under-vehicle setting. These observations should be interpreted as illustrative qualitative tendencies rather than as a quantitative benchmark.

4.2.4 Pipeline-Level Summary

For the implemented system, we therefore summarize the results at the pipeline level through visibility tolerance and component-wise latency scope rather than through stage-wise diagnostic tables. Any OLPD-related interpretation in this summary remains limited to the mask-given synthetic upper-bound under the oracle synthetic vehicle-mask setting; it does not evaluate automatic vehicle segmentation or deployment-time perception.

To make the MRV metric operational, let A_{vis} denote the number of visible pedestrian pixels in the occluded sample and A_{full} denote the number of pedestrian pixels in the paired non-occluded ground-truth mask. The visible ratio is defined as $r_{\mathrm{vis}} = A_{\mathrm{vis}}/A_{\mathrm{full}}$. We define Minimum Required Visibility (MRV) as the minimum visible ratio among test samples whose reconstruction satisfies both $\mathrm{IoU} \geq 0.80$ and $\mathrm{Dice\ (F1)} \geq 0.90$. Under this criterion, the observed MRV on the test set was 12.87%.

Because the original experiment log did not preserve the CPU/GPU model, inference runtime framework, batch size, or warm-up/repetition counts, Table 2 should be read as a local, hardware-unspecified latency summary of the measured reconstruction subpaths only. These values do not establish end-to-end real-time performance.

Table 2: Latency scope of the implemented reconstruction path.

Path	Included operation	Excluded operations	Mean latency
Keypoint filtering	post-OpenPose vehicle-mask-based joint filtering	OpenPose, vehicle-mask prediction, FBPI, PGSR, HSV	0.077 ms
FBPI forward pass	PartialToFullKeypointNet OpenPose, vehicle-mask prediction, keypoint filtering, PGSR, HSV		1.464 ms
PGSR forward pass plus resize/post-process	KeypointToMaskNet OpenPose, vehicle-mask prediction, keypoint filtering, FBPI, HSV		7.212 ms
FBPI+PGSR	FBPI inference + PGSR inference	OpenPose, vehicle-mask prediction, keypoint filtering, HSV	8.676 ms
Post-OpenPose path FBPI + PGSR	keypoint filtering + OpenPose, vehicle-mask prediction, HSV		8.753 ms

At the same time, the interpretation range of these results is narrow. The 10-background, 10-vehicle, and 10-child-pedestrian asset composition limits scene diversity, vehicle geometry variation, and especially the pose, scale, and silhouette distributions observed during training and evaluation. Because only child assets are used, the reported performance does not establish equivalent behavior for adult pedestrians or for broader variation in body shape, clothing, and self-occlusion. In addition, multi-person scenes and fully invisible cases are outside the evaluated distribution. Accordingly, the present numbers should be read as feasibility under controlled synthetic conditions rather than as evidence of robust real-world recovery.

5. Conclusions

5.1 Key Findings and Implications

This paper presented an implemented three-stage pose-guided reconstruction framework for partially visible pedestrians under vehicle occlusion. The current work focuses on recovering full-body structure and silhouette from sparse lower-limb evidence observable through the under-vehicle gap, rather than implementing a complete warning decision system. On the adopted synthetic dataset, the measured FBPI + PGSR reconstruction path required 8.676 ms per sample, excluding OpenPose BODY_25 inference, vehicle-mask prediction, keypoint filtering, and HSV, and adding only the post-OpenPose keypoint-filtering step yielded 8.753 ms. These values should be interpreted as local, hardware-unspecified measurements of the partial reconstruction path only, and they do not establish end-to-end real-time performance. Successful reconstruction was maintained down to a minimum visible ratio of 12.87% under the adopted criterion. The FBPI stage should be interpreted as full-body keypoint inference consistent with paired OpenPose reference poses rather than as measured human pose accuracy against manual keypoint annotation. In addition, PGSR in the current implementation is trained on paired OpenPose reference full-body poses from non-occluded images but is evaluated downstream with residual-fused FBPI outputs on occluded samples. Therefore, the reported PGSR behavior should be interpreted as a pose-conditioned upper-bound with respect to pose-input quality rather than as performance of a model trained on the exact deployment-time pose input. The additional oracle-versus-predicted mask comparison provides qualitative evidence only through an illustrative external segmentation example; it is not a reproducible benchmarked mask-prediction method specification in the current manuscript. At the same time, any OLPD-related interpretation in the current manuscript remains limited to the mask-given synthetic upper-bound under the oracle-mask setting and does not measure automatic vehicle segmentation or deployment-time perception. Taken together, these findings support feasibility under controlled synthetic conditions rather than robust real-world recovery, while HSV and downstream alert logic remain future work.

5.2 Limitations and Future Work

A notable limitation of this study is that the proposed Human Silhouette Verification (HSV) stage was not implemented or quantitatively evaluated in the current system. Although HSV was conceptually designed as a final quality-control module for rejecting implausible reconstructed masks, the present work reports no HSV-specific quantitative results and evaluates only the first three stages, namely lower-limb pose extraction, full-body pose inference, and pose-guided silhouette reconstruction.

A further limitation is that OLPD currently relies on the synthetic vehicle-layer mask available from the compositing pipeline. Accordingly, all reported OLPD results should be interpreted as mask-given synthetic upper-bound results for lower-limb keypoint filtering under a known occlusion mask. They do not evaluate automatic vehicle segmentation or deployment-time perception.

In addition, several aspects of the current experimental design directly limit how the reported results should be interpreted. The validation is restricted to a fully synthetic paired dataset constructed from 10 background images, 10 vehicle images, and 10 Pedestrian (Kid) assets. This small asset pool narrows scene diversity and constrains the pose, scale, and silhouette distributions seen during training and testing. Because the pedestrian assets are child-only, the reported numbers do not establish equivalent performance for adult pedestrians or for broader variation in body shape, clothing, and self-occlusion. Each sample also contains a single pedestrian instance, fully invisible cases were excluded because reliable lower-body keypoints could not be extracted by OpenPose, and the full-body keypoint supervision is derived from the paired OpenPose reference pose on the non-occluded image rather than from manual keypoint annotation. No manually annotated keypoint subset was used to validate actual human pose accuracy; therefore, FBPI metrics should be interpreted as consistency with OpenPose-derived reference poses. The same paired non-occluded reference pose is also used as the PGSR training input, whereas inference uses the residual-fused FBPI output on the occluded sample. Consequently, the present manuscript includes a pose-input train–test distribution gap at Stage 3, and PGSR results should be interpreted as pose-conditioned upper-bound evidence rather than as a fully matched train/test evaluation under predicted pose input. In addition, the surviving manuscript assets do not preserve the exact internal projection from the 25×3 PGSR pose tensor to the first spatial encoder feature map, so Fig. 6 should be interpreted as a conceptual schematic rather than a fully reproducible layer-by-layer architecture. Moreover, Fig. 9 uses an illustrative externally estimated mask example whose exact mask-prediction procedure is not specified as part of the benchmarked pipeline, so that panel should be interpreted as qualitative evidence only. Accordingly, the present quantitative results should be interpreted as a controlled feasibility study under restricted synthetic conditions rather than as evidence of broad road-scene generalization.

Future work should therefore include real-image validation, stricter cross-asset evaluation, adult pedestrians, broader body-shape and clothing variation, multi-person scenes, fully invisible cases, quantitative benchmarking with predicted masks including predicted vehicle-mask IoU, OLPD lower-limb F1 under predicted masks, FBPI OKS under predicted masks, PGSR IoU/Dice under predicted masks, separated mask-prediction latency, and implementation of the HSV stage.

References

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291–7299, 2017.

[2] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10863–10872, 2019.

[3] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "PARE: Part Attention Regressor for 3D Human Body Estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11127–11137, 2021.

[4] P. Sun, K. Gu, Y. Wang, L. Yang, and A. Yao, "Rethinking Visibility in Human Pose Estimation: Occluded Pose Reasoning via Transformers," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5903–5912, 2024.

[5] R. Khirodkar, S. Tripathi, and K. Kitani, "Occluded Human Mesh Recovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1715–1725, 2022.

[6] Q. Sun, Y. Wang, A. Zeng, W. Yin, C. Wei, W. Wang, H. Mei, C.-S. Leung, Z. Liu, L. Yang, and Z. Cai, "AiOS: All-in-One-Stage Expressive Human Pose and Shape Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1834–1843, 2024.

[7] Y. Zhu, A. Li, Y. Tang, W. Zhao, J. Zhou, and J. Lu, "DPMesh: Exploiting Diffusion Prior for Occluded Human

Mesh Recovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1101–1110, 2024.

[8] A. Stathopoulos, L. Han, and D. Metaxas, "Score-Guided Diffusion for 3D Human Recovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 906–915, 2024.

[9] B. Choi and S.-H. Lee, "K-SMPL: Korean Body Measurement Data Based Parametric Human Model," *Journal of the Korea Computer Graphics Society*, 28(4):1–11, 2022.

[10] D. Ahn and S. I. Park, "3D Human Shape Estimation from a Silhouette Image by using Statistical Human Shape Spaces," *Journal of the Korea Computer Graphics Society*, 29(1):13–22, 2023.

[11] H. Kim, K. Kang, G. Park, and T. Kwon, "Deep Learning-Based Motion Reconstruction Using Tracker Sensors," *Journal of the Korea Computer Graphics Society*, 29(5):11–20, 2023.

[12] J. Kim, D. Kang, Y. Lee, and T. Kwon, "Real-time Interactive Animation System for Low-Priced Motion Capture Sensors," *Journal of the Korea Computer Graphics Society*, 28(2):29–41, 2022.

[13] T. Zhang, B. Huang, and Y. Wang, "Object-Occluded Human Shape and Pose Estimation From a Single Color Image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7376–7385, 2020.

[14] Q. Zhou, S. Wang, Y. Wang, Z. Huang, and X. Wang, "Human De-Occlusion: Invisible Perception and Recovery for Humans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3691–3701, 2021.

[15] S. Xu, X. Li, and X. Lin, "2D Pose-guided Complete Silhouette Estimation of Human Body in Occlusion," in *26th International Conference on Pattern Recognition (ICPR)*, pp. 3618–3624, 2022.

[16] T. Xiang, A. Sun, J. Wu, E. Adeli, and L. Fei-Fei, "Rendering Humans from Object-Occluded Monocular Videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3239–3250, 2023.

[17] K. Xu, L. Zhang, and J. Shi, "Amodal Completion via Progressive Mixed Context Diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9099–9109, 2024.

[18] E. Ozguroglu, R. Liu, D. Suris, D. Chen, A. Dave, P. Tokmakov, and C. Vondrick, "pix2gestalt: Amodal Segmentation by Synthesizing Wholes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3931–3940, 2024.

[19] G. Zhan, C. Zheng, W. Xie, and A. Zisserman, "Amodal Ground Truth and Completion in the Wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28003–28013, 2024.

[20] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-Attention Mask Transformer for Universal Image Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299, 2022.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

[22] H. Fu, Y. Gu, Y. Yan, Y. Shen, Y. Wu, and L. Sun, "SDR-GAIN: A High Real-Time Occluded Pedestrian Pose Completion Method for Autonomous Driving," *arXiv preprint arXiv:2306.03538*, 2023.

[23] P. Hardy and H. Kim, "LInKs "Lifting Independent Keypoints" - Partial Pose Lifting for Occlusion Handling With Improved Accuracy in 2D-3D Human Pose Estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3426–3435, 2024.

[24] M. Armando, S. Galaaoui, F. Baradel, T. Lucas, V. Leroy, R. Bregier, P. Weinzaepfel, and G. Rogez, "Cross-view and Cross-pose Completion for 3D Human Understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1512–1523, 2024.

[25] H. Zheng, H. Li, W. Dai, Z. Zheng, C. Li, J. Zou, and H. Xiong, "HiPART: Hierarchical Pose AutoRegressive Transformer for Occluded 3D Human Pose Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16807–16817, 2025.

[26] W. Qu, J. Du, S. Yuan, J. Wang, Y. Sun, S. Liu, Y. Zhu, J. Rao, J. Yu, S. Cao, R. Xia, X. Tang, X. Wu, and D. Luo,

"DPGP: A Hybrid 2D-3D Dual Path Potential Ghost Probe Zone Prediction Framework for Safe Autonomous Driving," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 20341–20348, 2025.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, 2017.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

〈저자소개〉

유재천
- Current Position: Professor
- Affiliation: Department of Electrical and Computer Engineering, Sungkyunkwan University
- Education
- Ph.D. in Electrical and Electronic Engineering, POSTECH, 2001
- Career / Experience
- Senior Research Engineer, Samsung Electronics Co., Ltd., 1986 - 1996
- Professor, Samsung Institute of Technology, 1994
- Adjunct Professor, POSTECH Information Research Laboratories and Graduate School for Information Technology, 2002 - 2006
- Research Assistant Professor, POSTECH Graduate School for Information Technology, 2006 - 2008
- 관심분야: Design of deep learning architectures and their applications, AI-based medical image diagnosis, Design of AI-controlled robotic hands, AI-based digital healthcare platforms
- https://orcid.org/0000-0003-3721-0207

장현준
- Current Position: Master's Student
- Affiliation: Department of Electrical and Computer Engineering, Sungkyunkwan University
- Education
- B.S. in Electronic Engineering, Hankuk University of Foreign Studies, 2022
- M.S. Student in Department of Electrical and Computer Engineering, Sungkyunkwan University, 2025 - Present
- 관심분야: Image processing, Vision