Coming Out of the Dark: Human Pose Estimation in Low-light Conditions

Yong Su[1], **Defang Chen**[1], **Meng Xing**[2,3*], **Changjae Oh**[4], **Xuewei Liu**[5] and **Jieyang Li**[1]
[1]Tianjin Normal University
[2]University of Science and Technology of China
[3]Ningbo Institute of Digital Twin, Eastern Institute of Technology
[4]Queen Mary University of London
[5]Research & Development Branch, FAW Toyota Motor Co.,Ltd.
{suyong, xingmeng}@tju.edu.cn, chendefang@stu.tjnu.edu.cn, c.oh@qmul.ac.uk,
xw18002190986@163.com, ljieyang@foxmail.com

Abstract

Human pose estimation in low-light conditions is vital for applications such as surveillance and autonomous systems, yet the severe visual distortions hinder both manual annotation and estimation precision. Existing approaches typically rely on additional reference information to mitigate these issues, however, customized data collection equipment poses limitations on their scalability. To alleviate the issue, we construct a Low-Light Images and Poses (LLIP) dataset, which includes only paired low-light images and pose annotations obtained using off-the-shelf motion capture devices. Furthermore, we propose a Multi-grained High-frequency Feature Consistency Learning framework (MHFCL), which does not rely on additional reference information. MHFCL employs a Retinex-inspired restoration stream to recover high-frequency details and integrates them into pose estimation using a multi-grained consistency mechanism. Experiments demonstrate that our approach achieves a new benchmark in low-light pose estimation, while maintaining competitive performance in well-lit conditions.

1 Introduction

Human pose estimation (HPE) aims to locate the spatial coordinates of key human joints in an image by capturing multi-scale high-frequency information, making it a critical tool for a range of various downstream applications [Markovitz *et al.*, 2020]. While recent HPE models [Li *et al.*, 2021a; Wang *et al.*, 2020b] demonstrate strong performance in well-lit conditions, they struggle significantly in low-light conditions due to substantial image degradation. Some solutions use specialized sensors like LiDAR, depth, and thermal sensors to handle low-light environments. For instance, Lee et al. [Lee *et al.*, 2023] paired well-lit images with low-light ones from a dual-camera system and utilizes a teacher-student network to transfer privileged Information from well-lit image into low-light image. However, the collection and annotation of paired low-light and well-lit data is resource-intensive and error-prone, while the dependence on customized data collection equipments, which significantly increases both the training and deployment costs of low-light HPE models.

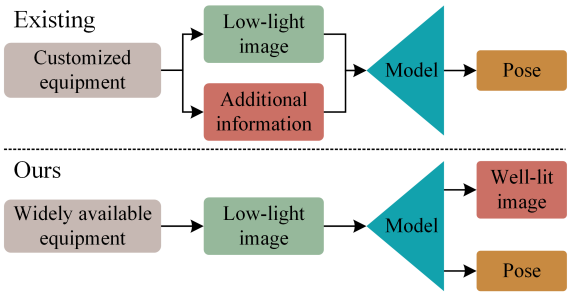

Figure 1: Comparative analysis of our approach against existing methods.

A feasible alternative is to collect only low-light images and pose annotations, then develop a model capable of estimating human pose in low-light conditions without relying on reference information, as shown in Figure 1. While this approach obviates the necessity for customized data collection equipments, it introduces two significant challenges: firstly, existing methods [Lee *et al.*, 2023] for low-light pose estimation typically rely on aligned well-lit images and pre-trained pose estimation models for annotations, yet the strategy is not applicable when only low-light images are available; secondly, in low-light conditions, the reduction of brightness and contrast amplifies the dominance of low-frequency components in the frequency spectrum, while simultaneously causing blurring and attenuation of high-frequency details. On the other hand, sensor noise, which is more pronounced in low-light conditions, manifests as randomly distributed high-frequency artifacts within the spectrum. In the absence of additional reference information, traditional human pose estimation (HPE) models encounter substantial difficulties in addressing the degradation of high-frequency information. This limitation impairs their capacity to learn robust feature representations, ultimately resulting in a marked decline in pose estimation precision [Wang *et al.*, 2020a].

*:Corresponding Author.

To address the aforementioned issues, we first present a Low-Light Images and Poses (LLIP) dataset, which provides paired low-light images and human poses annotations. Using off-the-shelf motion capture devices, we accurately recorded the 2D and 3D coordinates of 17 keypoints on the human body. Motion capture devices, unlike annotations based on pre-trained models or manual labeling, enable direct and accurate pose annotation, improving both estimation reliability and evaluation fairness. Moreover, LLIP provides both image and video with comprehensive annotations, paving the way for advanced research in low-light applications, such as action recognition and video anomaly detection in low-light conditions.

Then, we conceptualize reference-free low-light pose estimation within a two-stream architecture, termed Multi-grained High-frequency Feature Consistency Learning (MHFCL). The MHFCL framework integrates a Retinex-inspired High-frequency Restoration stream (RHR), which enhances pixel-level high-frequency components by jointly modeling global luminance distribution and local illumination dynamics. Notably, the RHR stream operates in an unsupervised manner, obviating the need for well-lit reference images as supervisory signals. The restored high-frequency components are subsequently fused into a Vision transformer-based Pose Estimation stream (VPE), enabling precise pose estimation under low-light conditions. To bridge the granularity gap between the two streams, we propose a multi-grained feature consistency learning mechanism that harmonizes pixel-level high-frequency feature maps with pose heatmaps, thereby enhancing feature coherence across the streams. Extensive experiments on low-light images validate the effectiveness of our approach, with the LLIP dataset and MHFCL being instrumental to its success. Furthermore, MHFCL demonstrates competitive performance in HPE tasks, underscoring its robustness and versatility.

The main contributions of this work are summarized as follows:

- A low-light human pose estimation dataset, LLIP, which consists of low-light images captured in various outdoor environments and their corresponding pose annotations, which demonstrates superior scalability by eliminating the necessity for customized data collection equipment.
- A multi-grained high-frequency feature consistency learning architecture, MHFCL, which restores the high-frequency degradation in an unsupervised manner and integrates it into the pose estimation process, thereby estimating pose in low-light conditions without additional reference information.
- Through extensive quantitative and qualitative evaluation, we demonstrate that our method shows competitive or state-of-the-art (SOTA) performance in two vision tasks, and analyses confirm that both our model and dataset play key roles in this success.

2 Related Works

2D Human Pose Estimation has advanced with deep learning, categorized into regression-based and heatmap-based methods [Zheng *et al.*, 2023]. Regression methods [Li *et al.*, 2023b; Panteleris and Argyros, 2022] focus on predicting joint coordinates from images. A pivotal advancement came from Toshev *et al.* [Toshev and Szegedy, 2014], who proposed DeepPose, a cascaded deep neural network regressor using AlexNet as the backbone to learn keypoints from images. After DeepPose, the research paradigm for HPE transitioned from classical methods to deep learning. The methods based on Res-Net backbone [Sun *et al.*, 2017], differentiable framework [Luvizon *et al.*, 2018], and cascade transformer [Li *et al.*, 2021b] have been proposed. Unlike directly estimating the 2D coordinates of human joints, heatmap-based methods in HPE aim to estimate 2D heatmaps by adding 2D Gaussian kernels at each joint position. Recent advancements include hourglass residual units (HRU) for multi-scale feature capture [Chu *et al.*, 2017], HRNet for accurate high-resolution keypoint predictions, GAN-based methods [Tian *et al.*, 2021] for biologically plausible poses generating, and Scale-Adaptive Heatmap Regression (SAHR) [Luo *et al.*, 2021] for scale adaptability improving. However, these methods depend on high-frequency information, and they struggle in low-light conditions, where texture and details are significantly degraded.

Downstream Low-light Vision Tasks. Recent advancements in vision algorithms have been successful under optimal weather and lighting conditions. However, their performance in challenging environments like fog, rain, snow, low light, and nighttime is limited. Such conditions significantly affect real-world applications, including autonomous vehicles, rescue robotics, and security systems, which require consistent functionality in adverse weather and lighting. Hong et al. [Hong *et al.*, 2021] proposed a robust object detection system for low-light conditions using a synthetic pipeline and recovery module. Chen et al. [Chen *et al.*, 2023] tackled instance segmentation in low light, employing adaptive weighted downsampling, smooth-oriented convolution blocks, and disturbance suppression. Lee et al. [Lee *et al.*, 2023] paired well-lit images with low-light ones from a dual-camera system and utilizes a teacher-student network to transfer privileged Information from well-lit image into low-light image.

Multi-grained Feature Extraction. Multi-grained features from CNNs have proven effective in various tasks. Li et al. [Li *et al.*, 2023a] explored multi-grained features from a transformer network for unsupervised Re-ID using a dual-branch architecture. Zhao et al. [Zhao *et al.*, 2021] proposed a "slow vs. fast" (SvF) learning strategy to balance old and new knowledge in few-shot class-incremental learning (FSCIL) with frequency-aware regularization. Zhang et al. [Zhang *et al.*, 2024] introduced a multi-grained spatio-temporal learning network for video anomaly detection (VAD), incorporating tasks like continuity judgment and missing frame estimation. Chen et al. [Chen *et al.*, 2021] proposed a multi-granular spatio-temporal graph network for skeleton-based action recognition, modeling both coarse and fine-grained motion patterns. Lastly et al. [Zhou *et al.*, 2022] designed a multi-granular self-supervised learning framework to enhance feature generality through instance and group discrimination.

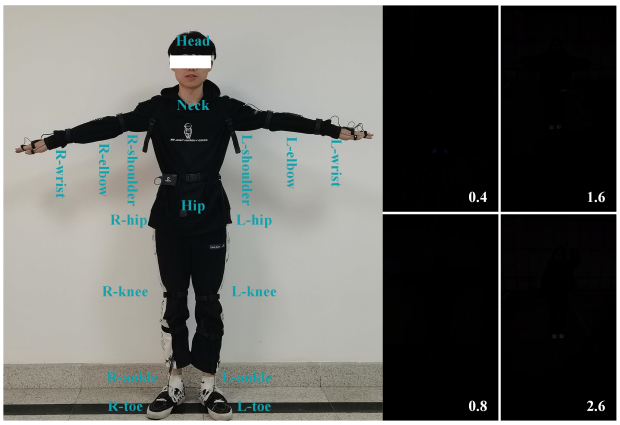

Figure 2: Overview of the LLIP dataset. Left: Configuration of motion capture devices and visualization of keypoints on the human body. Right: Visualization of low-light images under varying pixel intensity levels (the values in the bottom-right corner indicate the mean pixel intensity).

	Train/Test	Resolution	SC	MCA	SP
ExLPose	11405/2810	1920×1200	✗	✗	✗
LLIP	12378/4814	400×712	✓	✓	✓

Table 1: Comparison of LLIP and ExLPose in terms of Resolution, Standard Camera (SC), Motion Capture Annotation (MCA), and Spatial-temporal Pattern (SP).

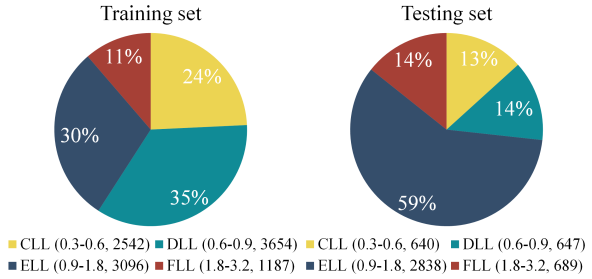

Figure 3: Distribution of training and testing samples by mean pixel intensity (range from 0.3 to 3.2) in the LLIP dataset.

3 LLIP Dataset

The LLIP dataset[1] consists of 12,378 low-light images for training and 4,814 images for testing. The training set includes images of eight individuals in two different scenes, while the test set spans five distinct scenes, allowing for an evaluation of model generalizability across varied conditions. The subjects, comprising five males and three females, range in height from 157 cm to 179 cm. All images are 400×712 pixels in resolution. The poses in the dataset are derived from randomly sampled asymmetric actions, with each action video beginning and ending in a T-pose. As illustrated in Figure 2, the images were captured in real-world low-light environments, and the corresponding pose information was concurrently recorded using motion capture devices. These devices capture data for 27 keypoints on the human body, from which the coordinates of 17 significant joints are extracted for pose estimation.

As outlined in Table 1, the LLIP dataset offers several advantages over existing low-light pose estimation datasets. Firstly, unlike many alternatives, LLIP does not require a custom-built dual-camera system, making the data collection process more cost-effective and replicable. Secondly, the use of motion capture devices for pose annotation eliminates the need for manual annotation or pre-trained pose estimation models. This not only ensures accurate pose labeling but also reduces the associated labor and computational costs. Additionally, the LLIP dataset is provided in both low-light image and video formats, making it suitable not only for low-light pose estimation but also for tasks such as low-light human action recognition and low-light video anomaly detection.

Lighting intensity is a critical factor influencing the degradation of high-frequency information and the amplification of noise in low-light images. To systematically investigate its effect on human pose estimation (HPE) performance, we categorized the images in the LLIP dataset based on their mean pixel intensity into four distinct groups: **C**hallenging **L**ow **L**ight (CLL), **D**ifficult **L**ow **L**ight (DLL), **E**xtremely **L**ow **L**ight (ELL), and **F**ormidably **L**ow **L**ight (FLL). This classification provides a structured framework for analyzing how varying levels of lighting intensity impact pose estimation accuracy. Figure 3 illustrates the distribution of training and testing samples across these categories, offering a detailed overview of the dataset composition. Notably, samples in the ELL and FLL categories account for over 70% of the testing set, making the LLIP dataset particularly challenging for low-light pose estimation tasks.

[1]The dataset is available on the project website: https://llip2024.github.io

4 Method

In this paper, we propose a novel reference-free framework for human pose estimation in low-light conditions, MHFCL. As illustrated in Figure 4, MHFCL incorporates two distinct streams: the RHR stream and the VPE stream. Inspired by the Retinex theory [Guo *et al.*, 2017], the RHR stream enhances the reflective component of the image in an unsupervised manner, utilizing reflectance & illumination decompose and a High-frequency aware loss to recover fine details obscured by noise. Meanwhile, the VPE stream generates keypoint heatmaps directly from low-light images. Furthermore, the Multi-grained Feature Consistency Learning mechanism (MFCL) combines pixel-level high-frequency features with pose heatmaps, thereby enhancing joint heatmap precision and overall pose estimation performance.

4.1 Retinex-based High-Frequency Restoration

The RHR stream consists of a series of convolutional layers, with strategically integrated Attention U-Net Block (AUB). RHR ultimately producing a 4-channel output tensor, where the first three channels correspond to reflectance (R) and the fourth to illumination (L), with intermediate layers using various convolution sizes and ReLU activations for channel transformations and feature fusion, as shown in Figure 4.

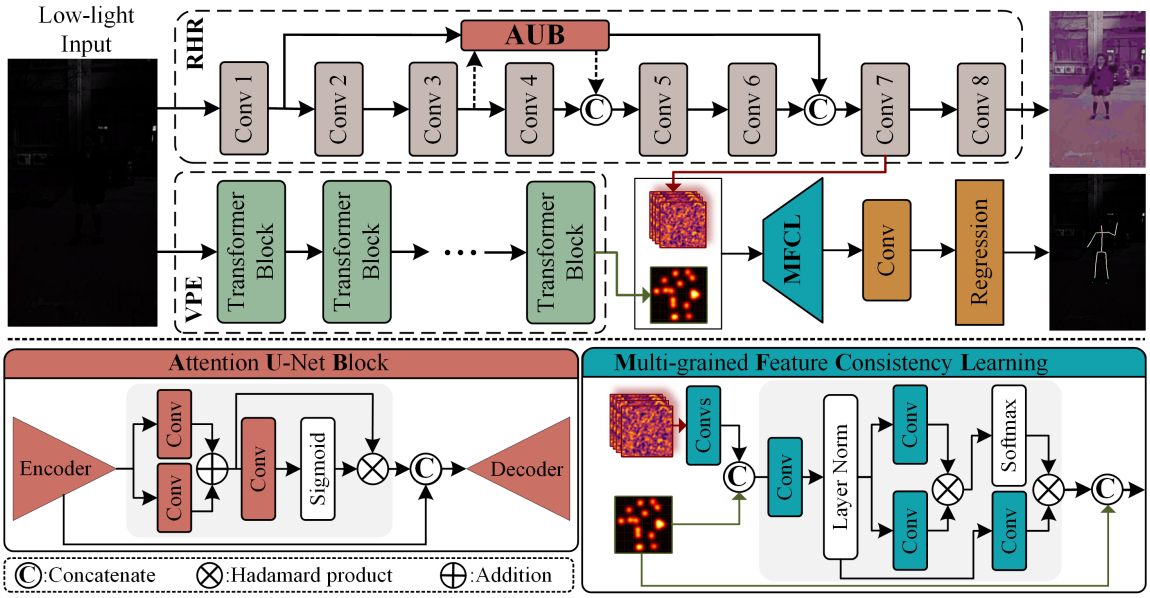

Figure 4: The architecture of the proposed Multi-grained High-frequency Feature Consistency Learning framework(MHFCL).

The AUB integrates attention mechanisms within the U-Net architecture to enhance the feature representation capability. Taking AUB between the first and seventh convolutional layers of the RHR stream as an example, the encoder consists of three convolutional layers. The attention mechanism is implemented with three convolutional layers. The first layer processes the input feature map. The second layer processes the gate signal and resizes it using bilinear interpolation. The outputs of these two layers are summed, and passed through a third convolutional layer with ReLU activation, generating an attention map, which is subsequently multiplied with the summed features to refine the final representation. Subsequently, the intermediate feature map of the encoder is also upsampled and concatenated with the illumination map. This concatenated feature is further processed through the decoder that consists of two convolutional layers followed by ReLU activations. We introduce three losses to enhance the recovery of fine details in low-light images, especially under noisy conditions.

The ℓ_1 is employed to mitigate high-frequency noise, especially from sharp image transitions or outliers, due to its robustness to such artifacts.

$$\ell_{recon} = ||R \cdot L - \mathrm{I}||_1 \tag{1}$$

The ℓ_{smooth} [Wei *et al.*, 2018] is designed to balance the preservation of structural boundaries while smoothing high-frequency texture features, effectively suppressing high-frequency noise.

$$\ell_{smooth} = ||I_x \cdot \exp(-\gamma \cdot R_x) + I_y \cdot \exp(-\gamma \cdot R_y)||_2 \tag{2}$$

where I_x and I_y represent the horizontal and vertical gradients of the input image I, R_x and R_y refers to the horizontal and vertical gradients of reflectance components R, the balancing coefficient γ is set to 10 to enhance structural awareness.

The ℓ_{const} [Guo *et al.*, 2020] ensures consistency across the color channels, thereby preventing color imbalances that could introduce abnormal high-frequency components.

$$\ell_{const} = (R_a - G_a)^2 + (R_a - B_a)^2 + (G_a - B_a)^2 \tag{3}$$

where R_{a}, G_{a}, and B_{a} denote the average values of the red, green, and blue channels.

4.2 Vision Transformer-based Pose Estimation

We developed a simple VPE stream based on the Vision Transformer (ViT) architecture [Dosovitskiy *et al.*, 2020], which serves as a stream to capture the structural and part high-frequency features of the human body. Following ViT [Dosovitskiy *et al.*, 2020], we divide each input image into fixed-size patches and convert each patch into a vector of specific dimensions. The transformer block constitutes the fundamental component of the ViT model, incorporating self-attention mechanisms and feedforward neural networks. In the self-attention mechanism, each patch is compared with other patches to compute their correlations and obtain attention-weighted patch representations, $\Theta_i, i \in \{1, \ldots, H\}$. Following [Xu *et al.*, 2022], we use the multi-head attention to represent multiple projections, where there are H different sets of learned projections instead of a single attention function. The final patch representation is derived by transforming the attention-weighted patch representation:

$$\hat{H}_n = \mathrm{MLP}(= \mathrm{Concat}(\Theta_1, \ldots, \Theta_H)W) \tag{4}$$

where W is the output projection matrix.

Given the ground truth heatmap $H_n \in \mathbb{R}^{192\times768}$ of a human pose, we write our heatmap loss for $\hat{H}_n$ as:

$$\ell_{pose} = \frac{1}{N}\sum_{n=1}^{N}\left\|\hat{H}_n - H_n\right\|^2 \tag{5}$$

where N refers the number of joints.

4.3 Multi-grained Feature Consistency Learning

The RHR stream operates at a pixel-level scale within low-level vision models, focusing primarily on recovering detailed information without explicit structural knowledge of the human body. In contrast, the VPE stream emphasizes both local and global body part structures, resulting in differences in granularity in their representations. To address this disparity, we propose a joint training approach that integrates the RHR stream with the VPE stream, introducing a multi-grained high-frequency feature consistency learning mechanism to ensure coherent feature representations across both networks. We apply three convolutional layers and bilinear interpolation operations to the output feature map $X \in \mathbb{R}^{64\times712\times40}$ from the 7-th layer of the RHR stream. The processed feature map $X_{\text{en}} \in \mathbb{R}^{192\times768}$ is then aligned with the heatmap $X_{\text{map}} \in \mathbb{R}^{192\times768}$ from the 11-th Transformer block of the VPE stream. Given the concatenated input features, we apply a convolutional layer with 1 channel to map and then normalize the output using layer normalization. Then we utilize self-attention to capture spatial correlations in the multi-grained feature map. Finally, we concatenate the restored high-frequency feature map with the original pose heatmap, resulting in $X = [X_{\text{map}}, X] \in \mathbb{R}^{2\times192\times768}$. After passing this fused feature map through a convolutional layer with 1 channel, we obtain the enhanced heatmap:

$$H = Conv(ReLU(X)) \tag{6}$$

Subsequently, we obtain the enhanced heatmap that can be dynamically optimized according to Eq. 5.

4.4 Loss Function

The total training loss ℓ_{total} is expressed as:

$$\ell_{total} = \lambda_1\ell_{pose} + \lambda_2\ell_{recon} + \lambda_3\ell_{smooth} + \lambda_4\ell_{const} \tag{7}$$

where λ_1, λ_2, λ_3, and λ_4 are the balance hyperparameters.

5 Experiment Results

5.1 Datasets

We conduct pose estimation experiments using three datasets: the proposed LLIP dataset, which includes images captured under low-light conditions, the ExLPose dataset, which comprises image pairs captured under low-light and well-lit conditions, and the widely used MS-COCO keypoint dataset, which comprises images captured under well-lit conditions. **ExLPose dataset [Lin *et al.*, 2014].** The ExLPose dataset contains 2,556 pairs of a low-light image and the corresponding well-lit image. The data is collected with a dedicated camera system and the pose annotations are generated by the pre-trained HPE model with the help of manual correction. **MS-COCO dataset [Lin *et al.*, 2014].** The MS-COCO dataset is a large-scale benchmark for human keypoint detection, comprising 150,000 labeled human instances for training, 5,000 images for validation, and 30,000 images for testing. It has become the most widely used benchmark for training and evaluating pose estimation models. We utilized the MS-COCO keypoints dataset to evaluate our method under well-lit conditions, comparing it with other baselines.

5.2 Evaluation Protocols and Parameter Settings

Our experiments were conducted on an RTX 3090 platform with a batch size of 6, utilizing the Adam optimizer. During the pre-training phase of the RHR stream, we augmented the LOL training set with an additional 303 images from the LLIP dataset, resulting in a total of 788 images. The initial learning rate was set to 1×10^{-4}. For the VPE stream, the learning rate was set to 2×10^{-4}. Both streams were pre-trained for 40 epochs. Subsequently, feature maps from the penultimate block of the RHR stream were extracted and used to enhance the heatmaps generated by the VPE stream. Joint training was then performed for 100 epochs. During the pre-training phase of the RHR stream, we applied regularization coefficients: $\lambda_1 = 0$, $\lambda_2 = 0.1$, $\lambda_3 = 1$, and $\lambda_4 = 0.5$. For the pre-training phase of the pose estimation network, we set $\lambda_1 = 1$ and $\lambda_2 = \lambda_3 = \lambda_4 = 0$. During the joint fine-tuning phase, regularization coefficients were adjusted to $\lambda_1 = 1$, $\lambda_2 = \lambda_3 = \lambda_4 = 0.1$.

To evaluate HPE performance on the MS-COCO dataset, we use the standard Object Keypoint Similarity (OKS) metric and report the Average Precision (AP). Without well-lit images, detection box annotation is challenging, as neither pre-trained detectors nor manual annotations can provide accurate boxes. To ensure accurate evaluation, we apply commonly used metrics on the LLIP dataset: Mean per Joint Position Error (MPJPE), Percentage of Correct Parts (PCP), Percentage of Correct Keypoints (PCK), and PCK_h, which provide a more reliable assessment in low-light conditions.

Method	Trained	Flops	MPJPE↓	PCK@10 ↑
ViTpose	✗	59.8G	90.3	24.4
ViTpose	✓	59.8G	13.8	44.6
Ours	✓	25.2G	**12.0**	**59.6**

Table 2: Comparison of pose estimation performance on LLIP dataset using various methods. "Trained" refers trained with LLIP.

5.3 Comparison with the State-of-the-art

Low-light HPE. Since the LLIP dataset does not include paired well-lit images, existing dual-stream low-light methods cannot be directly evaluated. Instead, we select ViTpose [Huang *et al.*, 2020], a model for well-lit conditions, as a baseline due to its strong performance and single-stream adaptability. Two experimental settings are considered: (1) directly testing the pre-trained ViTPose model on LLIP and (2) fine-tuning the pre-trained ViTPose model on LLIP before evaluation. As shown in Table 2, the pre-trained model struggles in low-light settings, while fine-tuning notably improves its performance. Our method further surpasses the fine-tuned ViTpose, demonstrating stronger robustness under low-light conditions. These results underscore the importance of the LLIP dataset for model adaptation and the effectiveness of our specialized approach in enhancing pose estimation accuracy.

Figures 5 and 7 compare our method with ViTPose on the LLIP and ExLPose datasets. On LLIP, where both models were trained, our method consistently outperforms ViTPose,

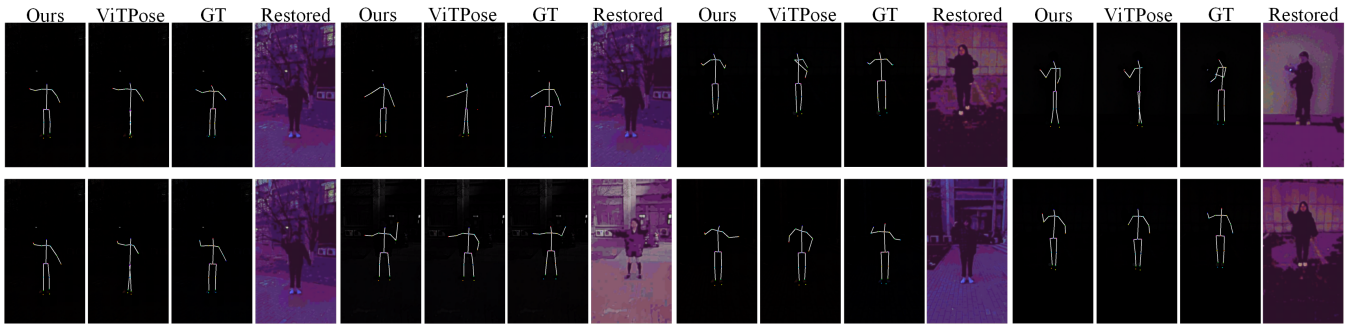

Figure 5: Qualitative results of our approach and comparison method on samples from LLIP dataset.

demonstrating superior precision. On ExLPose, where the pre-trained models were directly tested, our method achieves notably better results, showcasing its robustness in adapting to unseen datasets. These observations further emphasize the indispensable role of specialized low-light datasets and bespoke models in advancing low-light pose estimation.

Method	*Flops*	AP_{50}	AP_{75}
UDP [Huang *et al.*, 2020]	35.5G	**92.0**	84.3
TokenPose [Li *et al.*, 2021c]	22.1G	90.3	82.5
HigherHRNet [Cheng *et al.*, 2020]	154.3G	89.3	77.2
PRTR [Li *et al.*, 2021a]	21.6G	89.4	79.8
HRNetV1 [Wang *et al.*, 2020b]	7.1G	89.5	80.7
DarkPose [Zhang *et al.*, 2020]	32.9G	88.6	77.4
HRFormer [Yuan *et al.*, 2021]	29.1G	91.0	83.6
ViTPose [Xu *et al.*, 2022]	17.9G	90.7	83.2
ViTPose [Xu *et al.*, 2022]	59.8G	91.4	**85.2**
PCT [Geng *et al.*, 2023]	15.2G	91.2	84.7
Baseline	25.2G	90.8	86.3
Ours	25.2G	90.0	83.8

Table 3: Comparison of pose estimation performance on MS-COCO test set using various methods.

Well-lit HPE. To evaluate the model performance in well-lit conditions, we conduct experiments on the widely-used MS-COCO dataset. As summarized in Table 3, our baseline model, without RHR stream, achieves competitive performance with SOTA methods. However, incorporating the RHR stream leads to a slight performance decline, primarily due to the presence of overexposure and high-frequency noise in well-lit images. Although training on both low-light and well-lit images benefits performance in well-lit scenarios, it compromises robustness under low-light conditions. In contrast, our method preserves low-light robustness while ensuring reliable performance in well-lit environments.

5.4 Ablation Study

Effect of RHR stream. To assess the impact of the RHR stream on pose estimation, we conduct experiments under different mean pixel intensities (CLL, DLL, ELL, FLL). The results tabulated in Table 4 demonstrate that incorporating the RHR stream significantly improves pose estimation accuracy across varying pixel intensities. This enhancement is particularly evident under Formidably Low Light (FLL) conditions.

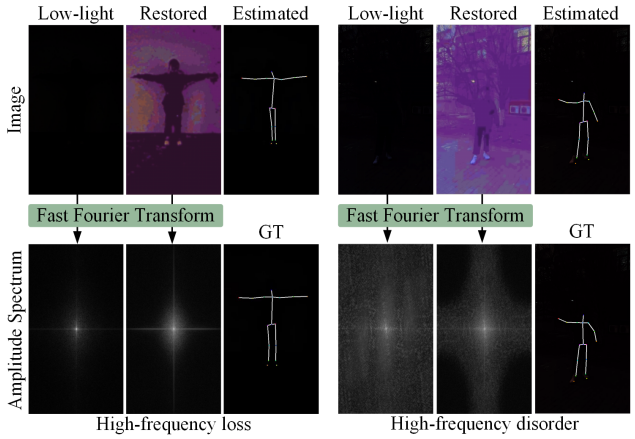

Figure 6: Two forms of high-frequency degradation in low-light conditions: high-frequency loss and high-frequency disorder.

As shown in the Figure 6, through sample spectrum analysis, we found that the RHR stream can effectively restore the loss of high-frequency information and the disorder of high-frequency information.

Performance on different mean pixel intensities. Pixel intensity in an image generally leads to a more pronounced degradation of high-frequency features and an increase in noise. However, contrary to intuition, Table 4 shows that weaker illumination does not necessarily result in larger errors in pose estimation. In low-light images, despite reduced illumination, higher ground reflectivity aids in distinguishing the ground from the human body, resulting in a higher PCK@10 value in FLL images compared to ELL. To further test this, we applied gamma correction to adjust the intensity of FLL images (Env_F) to match ELL (Env_E). As shown in Table 5, higher ground reflectance improves pose estimation under the same lighting conditions. This indicates that in low-light settings, ground reflectivity plays a critical role, unlike in well-lit environments.

Effect of VPE stream architecture variants. Table 6 shows that increasing the number of transformer blocks enhances pose estimation performance, with the model using 11 blocks achieving the highest accuracy. This indicates that deeper architectures offer improved representation power for

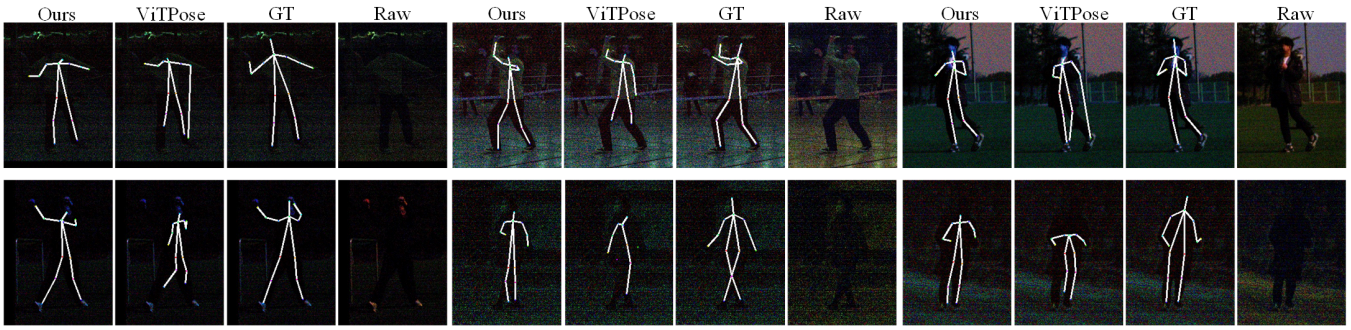

Figure 7: Qualitative results of our approach and comparison method on samples from ExLPose dataset.

Metrics	*Methods*	*CLL*	*DLL*	*ELL*	*FLL*
MPJPE ↓	*Baseline*	13.01	12.53	12.18	12.51
	Ours	12.38	11.89	12.24	**11.75**
PCK@10 ↑	*Baseline*	61.01	57.70	57.65	58.93
	Ours	**62.72**	59.42	57.07	59.91
PCK_h*@10*↑	*Baseline*	17.45	18.74	14.86	17.31
	Ours	18.26	17.69	13.30	**19.74**
PCK@50↑	*Baseline*	96.69	98.64	98.14	97.21
	Ours	96.87	**98.96**	98.11	97.88
PCK_h*@50*↑	*Baseline*	91.53	93.48	96.62	94.51
	Ours	92.34	95.35	**96.70**	95.67

Table 4: Performance across different mean pixel intensities on the LLIP dataset.

	MPGPE ↓	$PCK@10$ ↑	$PCK@50$ ↑
Env_F	**10.68**	**57.81**	**99.84**
Env_E	12.24	57.07	98.11

Table 5: Pose estimation results under low and high ground reflectivity conditions with identical pixel intensity.

pose estimation.

Metrics	5 Block	7 Block	9 Block	11 Block
MPJPE ↓	15.92	18.08	17.39	**12.00**
PCK@10 ↑	45.47	35.71	37.92	**59.64**
PCK@50↑	96.41	98.01	95.80	**98.40**

Table 6: Effect of the number of transformer block in VPE.

Effect of training paradigms. In addition to using end-to-end learning, we also test a two-stage training paradigm, where we first train the RHR stream and then use the high-frequency restored images to train the VPE stream. As tabulated in Table 7, the two-stage scheme achieves comparable performance with the end-to-end scheme in both well-lit and low-light conditions. However, in complex backgrounds, the performance of the two-stage scheme will significantly degrades as the severe interference of high-frequency information and noise in the background, as illustrated in Figure 8. In contrast, the end-to-end scheme prioritizes retaining the overall structure of the human body. Additionally, the inference time for the two-stage scheme per frame is 0.70 ms, whereas end-to-end is only 0.057 ms. Therefore, for all-weather HPE tasks, the end-to-end end-to-end is evidently superior to the two-stage method.

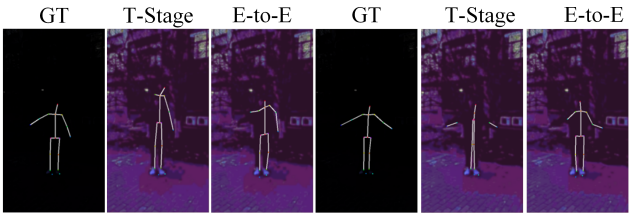

Figure 8: Qualitative results of challenging samples with different training paradigms.

well	E-to-E	T-Stage	low	E-to-E	T-Stage
MPJPE ↓	12.00	**11.81**	AP@50 ↑	**90.00**	87.50
PCK@10 ↑	59.64	**60.45**	AP@75 ↑	**83.80**	80.50
PCK@50 ↑	**98.40**	97.86			

Table 7: Effect of training paradigm. The well/low refers to the results in well-lit/low-light conditions. E-to-E and T-Stage refer to the end-to-end scheme and two-stage scheme.

6 Conclusion

We address the challenge of HPE in low-light conditions by introducing a new dataset, LLIP, and a novel reference-free HPE framework, MHFCL. MHFCL incorporates a RHR stream and a VPE stream, and enhances the strength of high-frequency features related to human body parts through a multi-grained feature consistency learning mechanism. Our results demonstrate that the proposed framework significantly enhances the accuracy of pose estimation in low-light conditions, without requiring well-lit images as reference information. Additionally, the proposed method achieves competitive results on well-lit images from the COCO dataset. This study demonstrates the efficacy of integrating high-frequency restoration with pose estimation to address the complexities of HPE under challenging lighting conditions. By eliminating the need for customized data collection equipments, our framework offers a valuable solution for practical applications.

References

[Chen *et al.*, 2021] Tailin Chen, Desen Zhou, Jian Wang, Shidong Wang, Yu Guan, Xuming He, and Errui Ding. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *Proceedings of the ACM international conference on multimedia, MM*, pages 4334–4342, 2021.

[Chen *et al.*, 2023] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023.

[Cheng *et al.*, 2020] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5386–5395, 2020.

[Chu *et al.*, 2017] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Geng *et al.*, 2023] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 660–671, 2023.

[Guo *et al.*, 2017] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017.

[Guo *et al.*, 2020] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1777–1786, 2020.

[Hong *et al.*, 2021] Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *Proceedings of the British Machine Vision Conference, BMVC*, volume 1, page 3, 2021.

[Huang *et al.*, 2020] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5700–5709, 2020.

[Lee *et al.*, 2023] Sohyun Lee, Jaesung Rim, Boseung Jeong, Geonu Kim, Byungju Woo, Haechan Lee, Sunghyun Cho, and Suha Kwak. Human pose estimation in extremely low-light conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 704–714, 2023.

[Li *et al.*, 2021a] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1944–1953, 2021.

[Li *et al.*, 2021b] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2021.

[Li *et al.*, 2021c] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE International conference on computer vision, ICCV*, pages 11313–11322, 2021.

[Li *et al.*, 2023a] Jiachen Li, Menglin Wang, and Xiaojin Gong. Transformer based multi-grained features for unsupervised person re-identification. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV*, pages 42–50, 2023.

[Li *et al.*, 2023b] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 17142–17151, 2023.

[Lin *et al.*, 2014] 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 *Proceedings of the European Conference on Computer Vision, ECCV*, pages 740–755, 2014.

[Luo *et al.*, 2021] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition, CVPR*, pages 13264–13273, 2021.

[Luvizon *et al.*, 2018] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018.

[Markovitz *et al.*, 2020] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 10539–10547, 2020.

[Panteleris and Argyros, 2022] Paschalis Panteleris and Antonis Argyros. Pe-former: Pose estimation transformer. In *International Conference on Pattern Recognition and Artificial Intelligence, ICPRAI*, pages 3–14, 2022.

[Sun *et al.*, 2017] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV*, 2017.

[Tian *et al.*, 2021] L. Tian, P. Wang, G. Liang, et al. An adversarial human pose estimation network injected with graph structure. *Pattern Recognition*, 115:107863, 2021.

[Toshev and Szegedy, 2014] 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, CVPR*, 2014.

[Wang *et al.*, 2020a] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 8681–8691. Computer Vision Foundation / IEEE, 2020.

[Wang *et al.*, 2020b] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[Wei *et al.*, 2018] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *Proceedings of the British Machine Vision Conference, BMVC*, page 155, 2018.

[Xu *et al.*, 2022] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems, NIPS*, 35:38571–38584, 2022.

[Yuan *et al.*, 2021] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. *arXiv preprint arXiv:2110.09408*, 2021.

[Zhang *et al.*, 2020] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7093–7102, 2020.

[Zhang *et al.*, 2024] Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 17385–17394, 2024.

[Zhao *et al.*, 2021] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow versus fast framework for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1576–1588, 2021.

[Zheng *et al.*, 2023] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023.

[Zhou *et al.*, 2022] Pan Zhou, Yichen Zhou, Chenyang Si, Weihao Yu, Teck Khim Ng, and Shuicheng Yan. Mugs: A multi-granular self-supervised learning framework. *arXiv preprint arXiv:2203.14415*, 2022.