

Bridging the Modality Gap: Advancing Multimodal Human Pose Estimation with Modality-Adaptive Pose Estimator and Novel Benchmark Datasets

Jiangnan Xia¹, Zhiyuan Zhang²[0000–0003–3945–5638], Yanyin Guo¹, Qilong Wu¹, Yi Li¹, Jianghan Cheng¹, and Junwei Li^{1*}

¹ College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China

{jiangnanxia, guoyanyin, 22231184, lee01, 22260382, lijunwei7788}@zju.edu.cn
<https://github.com/ICANDOALLTHINGS/Modality-Adaptive-Pose-Estimation>

² School of Computing and Information Systems, Singapore Management University, Singapore 178902, Singapore
cszyzhang@gmail.com

Abstract. Visual and infrared images represent two indispensable modalities that complement each other, offering unique insights into human pose estimation under different lighting conditions. However, existing efforts have predominantly focused on single modality, leading to significant challenges when transitioning to multimodal environments. The performance degradation observed in state-of-the-art models on multimodal images can be attributed to the substantial modality gap and the absence of multimodal benchmarks. To address this critical gap, we introduce novel visible-infrared multimodal human pose datasets where the two modality images are well balanced and accurately labeled. Leveraging these datasets, we establish the comprehensive benchmark to facilitate rigorous analysis and enhancement of multimodal human pose estimation techniques. Our findings underscore the limitations posed by modality variance on state-of-the-art methods. To overcome this challenge, we propose a method-agnostic scheme called Modality-Adaptive Pose Estimation, designed to seamlessly integrate into existing approaches. By employing Modality-Specific Batch Normalization and Modality Adaptive Loss, our approach enhances feature interactions between the two modalities, yielding superior performance. Extensive experiments conducted with popular baseline methods demonstrate the efficacy of our proposed approach in achieving state-of-the-art results on both modalities. We believe that our benchmarks offer a robust platform for investigating robustness and will significantly contribute to advancing research in this field.

Keywords: Human pose estimation · Multimodal · Visible · Infrared · Benchmark.

* corresponding author.

1 Introduction

Human Pose Estimation (HPE), a fundamental task in computer vision, holds significant importance across various domains, including human-computer interaction [44], motion capture [11], autonomous driving [69], and surveillance [9]. Current HPE approaches have primarily relied on leveraging single modality data, typically visual imagery thanks to the various available large-scale benchmarks [32, 2], with supreme performance achieved [4, 37, 31, 14, 54, 52, 21, 14]. However, the inherent limitations of single modal approaches become evident when confronted with low lighting conditions and complex weather.

To alleviate this challenge, several infrared HPE methods have proposed recently [34, 65, 70]. However, the number of infrared HPE methods is limited, and publicly available infrared benchmarks are scarce. Indeed, visual and infrared imaging modalities offer complementary information that can significantly enhance the robustness and accuracy of computer vision tasks. In recent years, we have witnessed that the integration of visual and infrared imaging modalities has emerged as a promising strategy in various applications such as human re-identification [23], object detection [5], image segmentation [42], Large Language Models (LLMs) [10], to name a few. However, in HPE, multimodal benchmarks are scarce mainly due to the difficulty of accurate labeling for both modalities. This hinders the application of HPE in real-world scenarios.

Addressing this critical gap necessitates the development of novel methodologies and comprehensive benchmark datasets that accurately represent the complexities of multimodal environments. The visible datasets are already rich and diverse [2, 32, 1, 33, 28], but collecting corresponding infrared images is a massive and tedious task. To resolve this challenge, we propose a novel method termed Instance Cross-modality Style Transfer Network(ICSTN) to intelligently transform visible images into infrared-style images, making full use of existing visible resources to construct diverse infrared datasets. Based on ICSTN, central to our contribution are three meticulously curated multimodal human pose datasets—COCO-MM and MPII-MM which are generated, and RAI-MM which is from the real world—where visual and infrared modality images are thoughtfully balanced and accurately labeled. Leveraging these datasets, we establish a comprehensive benchmark to facilitate rigorous analysis and enhancement of multimodal human pose estimation techniques.

Furthermore, we propose a method-agnostic scheme called Modality-Adaptive Pose Estimator(MAPE), designed to seamlessly integrate into existing methodologies. MAPE addresses the challenges posed by modality variance through the incorporation of Modality-Specific Batch Normalization(MSBN) and a novel Modality-Adaptive Loss(MAL), thereby enhancing feature interactions between visual and infrared modalities. Extensive experimental evaluations conducted with popular baseline methods demonstrate the efficacy of our proposed approach in achieving state-of-the-art results on multi-modalities.

In summary, our main contributions include:

- We propose an Instance Cross-modality Style Transfer Network(ICSTN), which enables the conversion from visible light images to infrared images.

This method optimizes instances with different heat levels in real infrared images, thus obtaining infrared images that are closer to reality.

- We construct multi-modality benchmarks COCO-MM, MPII-MM and RAI-MM, which include the original visible light data (COCO, MPII and RAI) and their corresponding infrared data (COCO-IRS, MPII-IRS and RAI-IR). Both parts of the data have supreme quality and precise annotations.
- We propose a novel Modality-Adaptive Pose Estimation(MAPE) method which can be easily applied to existing human pose estimation methods and exhibits strong performance simultaneously in both visible and infrared modalities. To the best of our knowledge, we are the first to address this challenging yet practical problem.

2 Related Works

Visual Image Human Pose Estimation. The field of human pose estimation (HPE) has been dominated by visible images. Since the pioneering work of DeepPose [56], numerous deep learning-based approaches have been proposed. Early works [6, 38, 53, 39, 30, 41] that usually regress the coordinates of key points directly for single-person pose estimation. However, the inherent difficulty of regression poses limitations on the accuracy and robustness of these models [13]. On the other hand, heatmap-based 2D pose estimation methods [46, 62, 4, 37, 31, 14] estimate per-pixel likelihoods for each keypoint location, and currently dominate in the field of 2D human pose estimation with more robust results achieved.

In recent years, there has been a growing focus on multi-person HPE, which presents a more realistic challenge. The methods can be broadly categorized into top-down and bottom-up approaches. The former [62, 51, 68] first detect individual humans and subsequently estimate pose key points within their bounding boxes. For example, SBL [62] employs a pedestrian detector and optical flow to determine detection boxes, integrating a deconvolution module for improvement. HRNet [51] distinguished by its parallel multi-resolution subnets, maintains high-resolution representations. On the contrary, bottom-up methods [26, 27, 24, 61, 47] independently detect key points for all individuals and then employ association strategies to combine them into complete skeletons for each person. Notably, HigherHRNet [8] advances bottom-up pose estimation by generating multi-resolution heatmaps to handle scale variation. Further enhancements in performance have been achieved through techniques such as embedding association [45], optimizing the heatmaps [47], or regressions [17]. More recently, with the introduction of new techniques such as Transformer [58] and Diffusion [20], the performance of both bottom-up and top-down methods has been further elevated to a new level [54, 52, 21, 14]. However, achieving these outstanding results often requires adequate lighting, which is not always available in real-world scenarios.

Infrared Image Human Pose Estimation. Despite the significant progress on visible image HPE, the performance degrades when lighting becomes dark.

To tackle this issue, infrared image HPE also attracts much attention in recent years. Currently, there is limited research in this field, and most studies [34, 65, 70] have drawn inspiration from visible image HPE. Liu et al. [34] proposed elliptical distribution encoding learning for human pose regression and constructed anisotropic Gaussian label for adjacent limbs connection under infrared imaging. Zhu et al [70] proposed a novel model called FEPose which incorporates the Transformer Encoder architecture and a specially designed FELayer layer for infrared images to improve the accuracy of human pose estimation. Xu et al. [65] proposed InfPose, an infrared multi-human pose estimation based on a lightweight encoder-decoder CNN for edge devices with a wild infrared human pose dataset established. While these approaches demonstrate promising performance in infrared image HPE, the lack of a unified methodology effective for both visible and infrared domains remains a significant research challenge, which is the primary focus of this paper.

Human Pose Estimation Benckmarks There have been numerous well-established benchmarks in the realm of visible image HPE. benchmarks like LSP [25] and FLIC [49] were pivotal in the early stages of single-person HPE; however, their performance is constrained by data scale limitations. The advent of deep learning-based methodologies has ushered in a proliferation of benchmarks [2, 32, 1, 33] comprising large-scale datasets. For instance, MPII [2] aggregates human images from media videos, culminating in a vast and diverse HPE dataset encompassing 40,522 annotated images of individuals, complete with coordinates for 16 joints. Another popular HPE benchmark is COCO [32], with 250,000 labeled individuals, caters to a broad spectrum of everyday life scenarios, utilizing metrics such as AP, AR, and their variants for evaluation purposes. Additionally, HiEve [33] introduces more challenging scenarios like crowded scenes and earthquake evacuation, and pioneers the use of the weighted AP (w-AP) metric to incentivize model performance in complex scenes. In contrast, benchmarks for infrared image HPE are limited and not readily available to the public. The benchmark closest to addressing this gap is SLP [36], a multimodal large-scale lying pose dataset that includes RGB and infrared modalities. This dataset, obtained through physical hyperparameter tuning, primarily serves the application of in-bed human pose monitoring. To address this gap, we propose a novel large-scale multimodal HPE benchmark based on MPII and COCO datasets that can be used for diverse real-world scenarios.

3 Methods

3.1 Instance Cross-modality Style Transfer Network

Our first objective is to establish a large-scale multimodal human pose benchmark. To fully utilize the existing visual image data source on HPE, we propose to generate realistic infrared human pose images from the visual image datasets.

Compared to visible images, infrared images place greater emphasis on instance-specific information[43]. In HPE, individuals display discrepancies in infrared intensity compared to other elements, leading to fluctuating brightness levels across

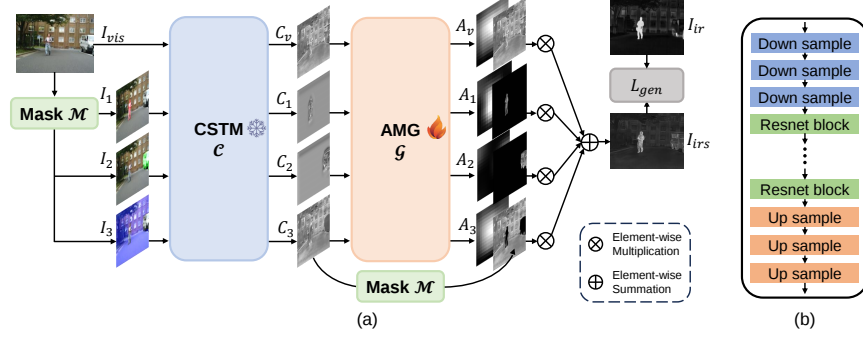


Fig. 1. (a) The framework of the ICSTN. (b) The structure of CSTM and AMG.

different instances in infrared images [57]. However, existing pseudo-infrared generation methods, such as those utilized for image registration tasks [12], predominantly emphasize structural characteristics, overlooking instance-specific variations, and generating less realistic images.

Based on this observation, we propose an Instance Cross-modality Style Transfer Network (ICSTN) comprising two stages to transform visible images I_{vis} into infrared-style images I_{irs} in Figure 1(a). In the first stage, we use a segmentation network $\mathcal{M}()$ [18] to isolate instances in the image based on their different thermal energy in the real world. In our implementation, we have defaulted to three categories: high thermal energy instances (e.g., humans, animals), medium thermal energy instances (e.g., cars, motorcycles), and low thermal energy instances (e.g., plants, bicycles). Therefore, the segmentation results I_i are related to the thermal energy levels. Then, a pre-trained Cross-modality Style Transfer Module (CSTM) $\mathcal{C}()$ is used to directly convert the original visible image I_{vis} and segmented visible image I_i into the corresponding infrared-style image C_v and C_i . In the second stage, to obtain more realistic infrared images, we propose Attention Map Generator (AMG) $\mathcal{G}()$ based on the UNet [48] architecture to generate attention maps A_v and A_i for C_v and C_i which are then used as weights for fusion. Here, we use the style loss [16] as the training loss L_{gen} forcing the AMG to generate attention maps to fuse more realistic infrared images. Finally, the infrared-style images I_{irs} can be obtained via Equation 1:

$$I_{irs} = A_v \otimes C_v + \sum_{i=1}^K A_i \otimes \mathcal{M}(C_i), \quad (1)$$

where $\otimes$ is the Hadamard product and K indicates the number of thermal energy instances which is 3 throughout this paper.

The architecture of CSTM is detailed in Figure 1(b). CSTM has 9 ResNet [19] blocks, 3 convolution layers for downsampling, and 3 transposed convolution layers for upsampling. We follow [12] and pre-train CSTM on the RoadScene dataset [64]. During the training of ICSTN on MSRS[55], we freeze the parameters of CSTM and update the parameters of AMG. This ensures that while

maintaining the style transfer capability of CSTM, AMG can generate attention maps to produce infrared images that more close to real ones. The AMG shares the same architecture as the CSTM but includes 6 ResNet blocks.

3.2 Multi-modality Pose Benchmark

Benchmark datasets Based on the proposed ICSTN, we are now able to construct large-scale benchmark datasets, termed COCO-MM and MPII-MM, respectively. Furthermore, we propose a novel RAI-MM, containing real-world visible dataset RAI and infrared dataset RAI-IR. The datasets proposed provide the previously data-scarce field with ample and sufficiently diverse datasets, ensuring fairness and openness in research.

The COCO-MM is constructed on the train2017 set and val2017 set of the COCO [32]. The entire dataset comprises the original dataset COCO and its corresponding infrared-style dataset COCO-IRS, which aligns well with the COCO dataset and utilizes the same annotations. A partial display of COCO-MM is shown in Figure 2.

The MPII-MM is built upon the MPII [2], comprising the original dataset MPII and its corresponding infrared-style dataset MPII-IRS, which aligns well with the MPII dataset and utilizes the same annotations. A partial display of MPII-MM is shown in Figure 3.

The RAI-MM is proposed as a valid benchmark to assess multimodal HPE in real-world scenarios. We select 600 pairs of images from *MSRS* [55] and *M³FD* [35] and capture 2400 pairs of visible and infrared images using well-matched visible-infrared cameras. The whole dataset consists of 3000 pairs of visible-infrared images and 18420 human instances, which consists of the visible dataset RAI(3000 visible images and 9150 human instances) and the infrared dataset RAI-IR(3000 infrared images and 9270 human instances). The RAI-MM is divided into the well-light part and the low-light part. Each part is organized into indoor and outdoor scenarios, with the outdoor scenarios further divided into road scenes and campus scenes. Figure 4(a) and (b) show the distribution of the dataset. We follow the COCO [32] and annotate each modality employing 17 key points and ground truth bounding boxes, with details shown in Figure 4(c). It is worth noting that the visible images in RAI-MM include both well-light scenes and low-light scenes, which is beneficial for tests under various lighting conditions. A partial display of RAI-MM is shown in Figure 5. Moreover, we also use ICSTN to generate RAI-IRS based on RAI, which is used to test the generation quality of ICSTN in Sec 4.1.

Evaluation Metrics For the COCO-MM dataset, the official COCO evaluation metrics Average Precision(*AP*) and Average Recall(*AR*) are used to assess the model’s performance. Moreover, we introduced *mmAP* and *mmAR* to evaluate the overall performance in multimodal scenarios.

$$mmAP = \frac{1}{N_m} \sum_{m=1}^{N_m} AP_m, \quad mmAR = \frac{1}{N_m} \sum_{m=1}^{N_m} AR_m, \quad (2)$$



Fig. 2. Examples from the COCO-MM dataset. The first four rows depict examples of single-person poses, while the last four rows depict examples of multi-person poses.

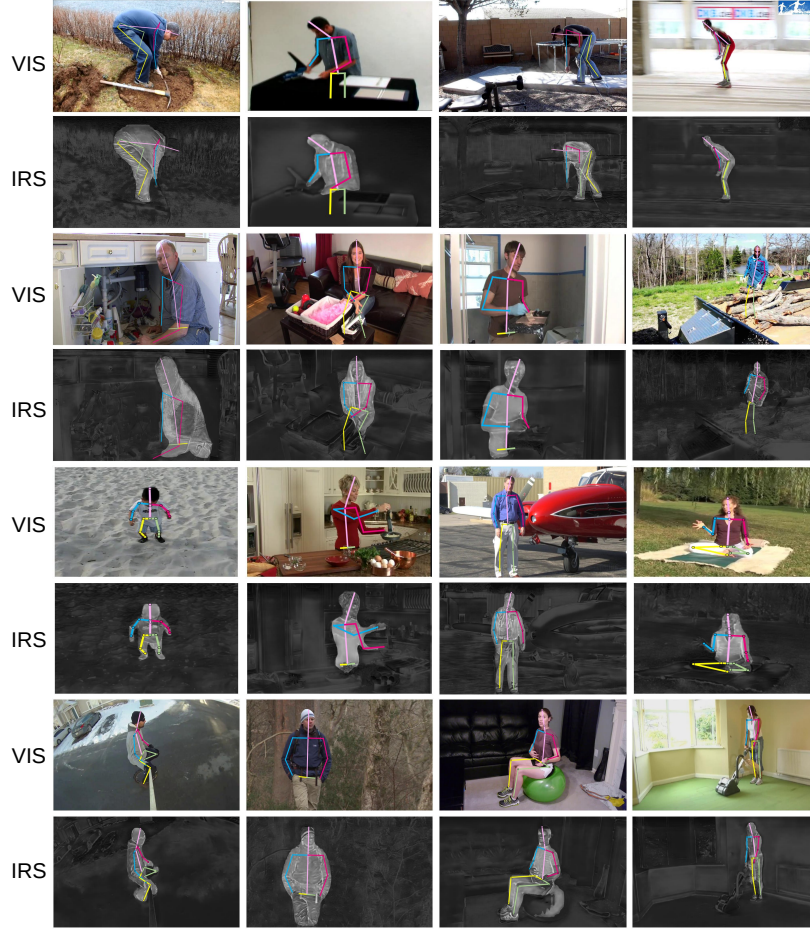


Fig. 3. Examples from the MPII-MM dataset.

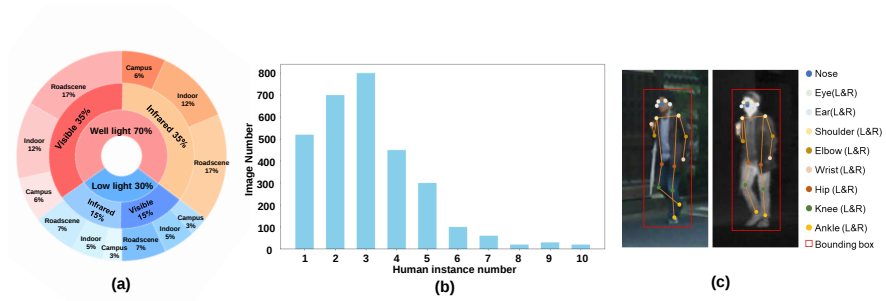


Fig. 4. Details of RAI-MM. (a) Distribution of scenes; (b) Distribution of the number of human instances; (c) Definition for annotations (L: left, R: right).

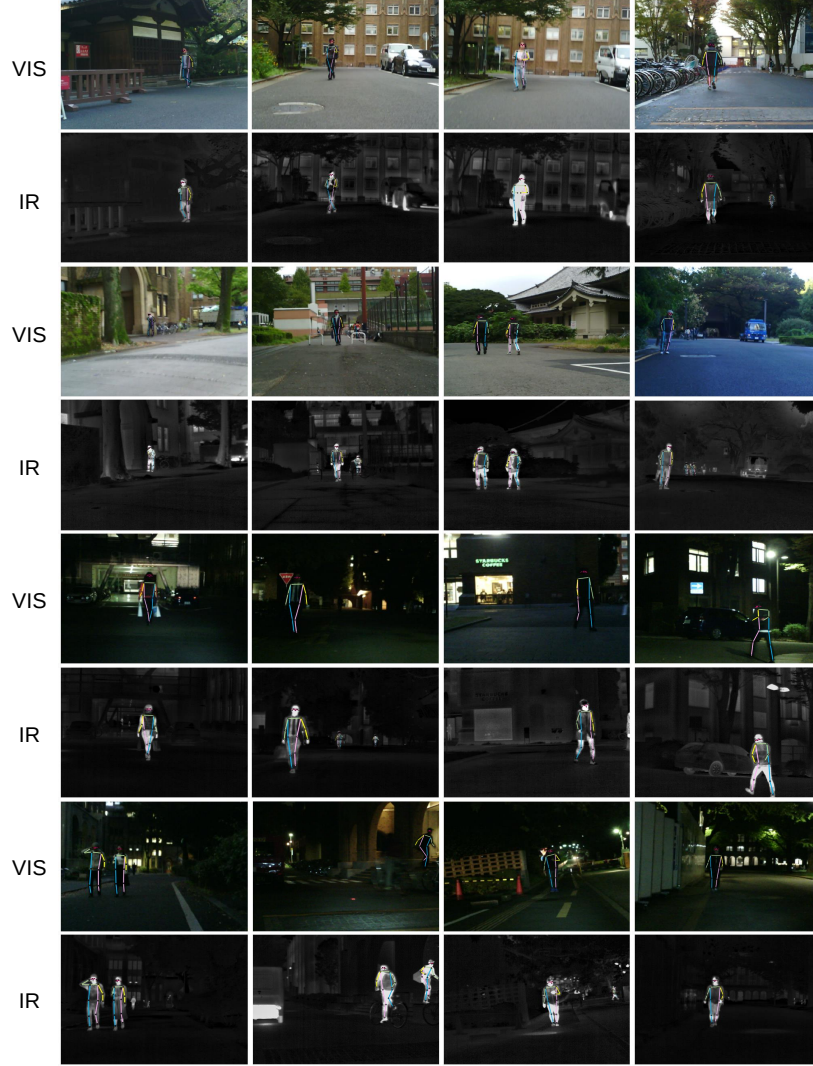


Fig. 5. Examples from the RAI-MM dataset. The first four rows depict examples of well-light scenes, while the last four rows depict examples of low-light scenes.

where AP_m and AR_m represent the AP and AR in a specific modality. In this work, $N_m = 2$ indicates the two modalities. $AP_m \in \{AP_v, AP_{ir}\}$, $AR_m \in \{AR_v, AR_{ir}\}$, where AP_v and AR_v denote the accuracy on the visible modality, while AP_{ir} and AR_{ir} denote the accuracy on the infrared modality.

For the MPII-MM dataset, the official MPII evaluation metric $PCKh$ is used. Similar to Equation 2 we define $mmPCKh$ as follows (Equation 3) to integrate the performance from different modalities.

$$mmPCKh = \frac{1}{N_m} \sum_{m=1}^{N_m} PCKh_m. \quad (3)$$

For the RAI-MM dataset, the same evaluation metrics as COCO-MM are used to assess the model’s performance. Similarly, we used $mmAP$ and $mmAR$ to evaluate the model’s comprehensive performance on the multi-modality benchmark RAI-MM. AP_v and AR_v denote the accuracy on the visible modality benchmark RAI, AP_{ir} and AR_{ir} denote the accuracy on the infrared modality benchmark RAI-IR.

3.3 Modality-Adaptive Pose Estimation

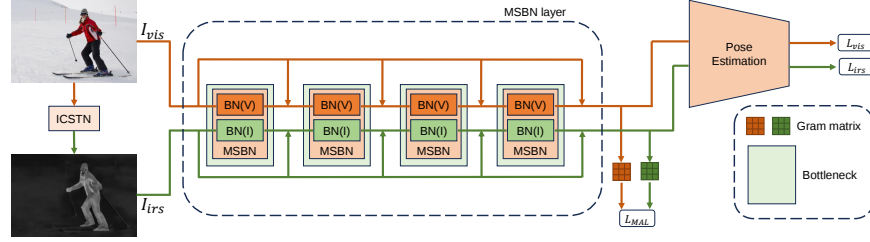


Fig. 6. Illustration of the detailed MAPE architecture.

Current Human Pose Estimation (HPE) methods predominantly focus on single modality data, limiting their applicability across diverse modalities. To address this limitation and achieve precise HPE for both visual and infrared modalities, we introduce the Modality-Adaptive Pose Estimation (MAPE) method. The architectural overview of MAPE is depicted in Figure 6. The method is designed to leverage distinct batch normalization techniques for each modality, while sharing the remaining network parameters. Drawing inspiration from domain-specific batch normalization [7], we propose Modality-Specific Batch Normalization (MSBN) to capture modality-specific characteristics for both visible and infrared modalities. By integrating MSBN, we can mitigate modality-specific biases during training, facilitating the network in capturing modality-invariant features more effectively. Moreover, to enhance the network’s adaptability to modality variations, we introduce a novel loss function termed Modality-Adaptive Loss

(MAL) (see section 3.3). By integration of MSBN and MAL, MAPE can be easily applied to existing models and achieves robust HPE in multi-modal scenarios.

Modality-Specific Batch Normalization. In the evaluated benchmarks, individual poses are considered to embody modality-invariant information, whereas distinct modality styles (e.g., visible and infrared) encapsulate modality-specific characteristics. Hence, for training on multi-modal datasets, we advocate the use of paired visible images and infrared-style images to facilitate the learning of modality-specific information. During each training epoch, both visible images and their corresponding infrared-style counterparts are jointly utilized as inputs. Leveraging different batch normalizations within Modality-Specific Batch Normalization (MSBN) based on their respective modalities enables effective elimination of modality-specific biases. This approach assists the network in fostering modality-invariant learning, thereby enhancing its adaptability across diverse modalities and ensuring robust performance in multi-modal scenarios.

Let $x_{mod} \in \mathbb{R}^{N \times H \times W}$ denote activation at each channel belonging to a modality label $mod \in \{vis, irs\}$, the MSBN can be expressed as:

$$MSBN_{mod}(x_{mod}; \gamma_{mod}, \beta_{mod}) = \gamma_{mod} \cdot \hat{x}_{mod} + \beta_{mod}, \quad (4)$$

where $\hat{x}_{mod} = (x_{mod} - \mu_{mod}) / (\sqrt{\sigma_{mod}^2 + \epsilon})$. Here, μ_{mod} and σ_{mod}^2 denote the mean and variance of the activation in x_{mod} , respectively, γ_{mod} and β_{mod} are affine transform parameters in batch normalization for the specific modality, and ϵ is a small constant used to prevent division by zero.

Referring to the stem layer of HRNet [51], we use four bottlenecks to construct the MSBN layer, in which batch normalizations [22] are replaced by MSBN blocks. MSBN layer can effectively reduce the impact of modality differences on the model. Its premise lies in the accurate encoding of different modality images, which means that different modality features need to be distinguished in the shallow layers of the network. We use t-SNE [40] to visualize the encoding of different modalities in the shallow layers of HPE network. As shown in Figure 7, it can be observed that using MSBN enables accurate differentiation between different modalities, while without using MSBN, modality mixing may occur.

Modality-Adaptive Loss To enable effective information exchange between the visible and infrared modalities, we introduce a novel loss function termed Modality-Adaptive Loss(MAL). Unlike the conventional adaptation from source domain to target domain [15, 59], visible and infrared are two complementary modalities. Complementary information needs to be fully utilized. Therefore, MAL is designed to perform feature interaction and alignment for the two modalities. Operating concurrently on both visible and infrared branches, MAL facilitates alignment of modality-specific features generated by MSBN. By minimizing feature discrepancies at the input of subsequent pose estimators, MAL encourages the model to adapt to modality variations. Consequently, the model learns

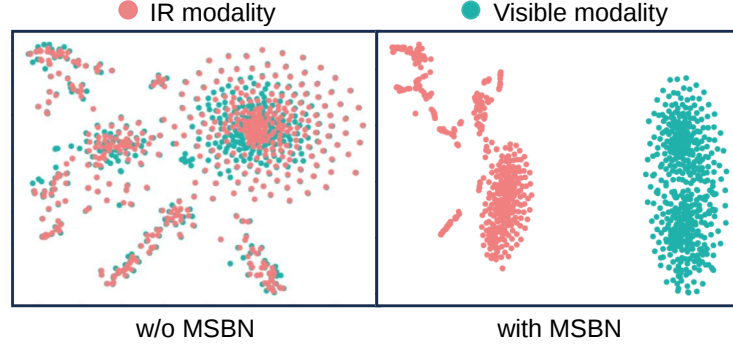


Fig. 7. t-SNE visualization of the encoding of different modalities in the shallow layers of the network.

representations that are invariant to modality enhancing its capability to generalize across different modalities and ensuring robust performance in multi-modal scenarios.

We utilize the Gram matrix [16] to compute the modality style according to the modality-specific feature F^{mod} processed by MSBN. Let $G^{mod} \in \mathbb{R}^{C \times C}$ represent the feature correlations between C channels' feature maps of F^{mod} in mod modality. It can be calculated by:

$$G_{ij}^{mod} = \sum_k f_{ik}^{mod} f_{jk}^{mod}, \quad (5)$$

where f_{ik}^{mod} and f_{jk}^{mod} are activations from the i^{th} channel and j^{th} channel of F^{mod} at position k , respectively. In our paper, $mod \in \{vis, irs\}$. Then our MAL can be computed as:

$$L_{MAL} = \frac{1}{4C^2M^2} \sum_{i,j=1}^C (G_{ij}^{irs} - G_{ij}^{vis})^2, \quad (6)$$

where C and M are the number of channels and the spatial size of the feature F , respectively. By minimizing the above loss function, the gap between modalities is reduced, aiding the model in achieving modality adaptation.

4 Evaluation on Multi-modality Pose Benchmark

4.1 Cross-modality Style Transfer analysis

We compare different state-of-the-art cross-modality style transfer methods for transforming visible images into infrared-style images. As illustrated in Figure 8, the models used are all trained on RoadScene [64] and MSRS [55]. When using

GPTN [3], the generated pseudo infrared image suffers great structural degradation. When using CPSTN [12], the model retains structural information but fails to perform modality conversion effectively, missing highlighting the brightness differences that infrared images typically exhibit due to inconsistencies in heat distribution. In contrast, our model achieves successful modality conversion while retaining a certain degree of structural information.



Fig. 8. Comparisons between existing cross-modality style transfer methods and our proposed method on the RAI-MM dataset. It can be observed that our method generates infrared-style images that are closer to real infrared images.

Figure 9 further shows visual comparisons of features extracted at different stages in the HPE model, which shows that the differences between IR and IRS are minimal, providing further evidence that the generated infrared images are trustworthy in HPE.

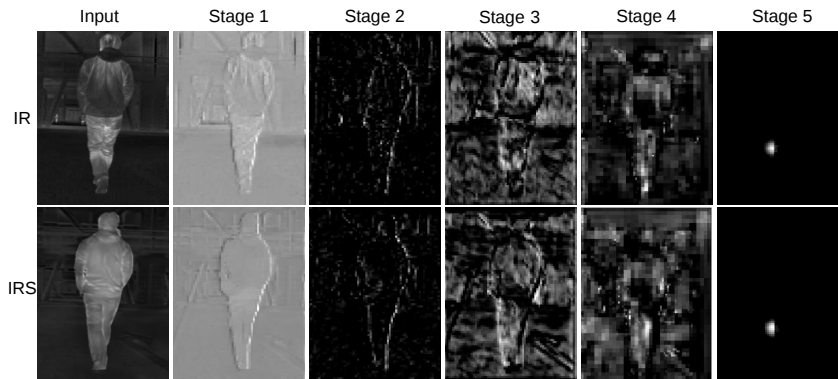


Fig. 9. Comparisons between IR and IRS in different HPE stages.

To quantitatively analyze the generated infrared images, we employ the structural similarity index (SSIM) metric and the peak signal-to-noise ratio (PSNR) metric. By computing the SSIM and PSNR values between the infrared-style images generated from visible images and their corresponding ground truth infrared images, we obtain the quantitative results shown in Table 1. It can be observed that our method can generate images that are closer to real infrared images. To better validate the quality of generated infrared images compared to the real infrared images, we train models on COCO-IRS and test them on RAI-IR and RAI-IRS (generated by ICSTN based on RAI) in Table 2, which shows close performance for models between IR and IRS, demonstrating the reliable quality of the generated dataset in real infrared scenarios validation.

Table 1. Comparisons between the existing method and our proposed method on the RAI-MM dataset. Our method outperforms the existing method in SSIM and PSNR.

Method	Backbone	SSIM $\uparrow$	PSNR $\uparrow$
GPTN [3]	CycleGAN	16.7	5.2
CPSTN [12]	CycleGAN	18.9	6.8
ICSTN(Ours)	CycleGAN	68.9	20.4

Table 2. Performance of HPE methods on RAI-IR and RAI-IRS. AP_{ir} means performance on RAI-IR, while AP_{irs} means performance on RAI-IRS.

Method	Backbone	Input size	AP_{ir}	AP_{irs}
SBL [63]	ResNet101	256×192	66.5	66.8
HRNet [51]	HRNet-W32	256×192	68.9	69.2
HRNet [51]	HRNet-W48	256×192	69.7	69.9
HrHRNet [8]	HrHRNet-W48	512×512	58.1	58.3
ViTPose-B [66]	ViT-B	256×192	72.0	72.3
FEPose-B [70]	FEPose-B	256×192	71.9	72.1
InfPose [65]	InfPose	512×512	60.0	60.3

We further conduct an ablation study on the thermal instance K , and the results are shown in Table 3. Specifically, 1 represents distinguishing only high-thermal instances, 2 represents distinguishing high-thermal and medium-thermal instances, and 3 represents distinguishing high-thermal, medium-thermal, and low-thermal instances. The results indicate that using all three thermal instances leads to better performance.

4.2 Experimental Setup

We assess the performance of state-of-the-art visible HPE methods [63, 51, 8, 67, 66] and infrared HPE methods [70, 65] on the proposed multi-modality pose

Table 3. Ablation study on the number of thermal energy instances.

K	SSIM $\uparrow$	PSNR $\uparrow$
1	66.5	17.8
2	67.8	19.1
3(Ours)	68.9	20.4

benchmark RAI-MM. To evaluate the capability of models trained on single modality when facing multi-modality tasks, we follow the official settings of each method and train them separately on COCO and COCO-IRS. Moreover, to validate the effectiveness of the proposed multi-modality datasets, we then train methods on COCO-MM. The results are presented in Table 4.

4.3 Benchmark Conclusions

The results shown in Table 4 indicate that models trained on the single modality significantly drop in performance when applied to multi-modality scenarios, whereas models trained on multi-modality datasets exhibit consistent and satisfactory performance. These findings underscore the critical role of the multi-modality dataset proposed in this study. It is also worth noting that directly introducing multimodal datasets to improve overall multimodal performance can somewhat compromise the performance of individual modalities. This underscores the necessity of our proposed method.

Furthermore, we analyze the factors influencing the model’s performance in multi-modality applications. As shown in Table 4, under the same HPE method, the performance varies with changes in the backbone and input resolution. We observe that when input resolution is consistent, using a backbone with larger capacity (HRNet-W48) leads to higher accuracy and modality robustness compared to using a backbone with lower capacity (HRNet-W32). Similarly, when the backbone is consistent, using a larger input resolution (384×288) results in higher accuracy and modality robustness compared to using a lower input resolution (256×192). This suggests that stronger backbones and larger input resolutions can enhance the model’s multi-modality performance.

5 Multi-modality Pose Estimation with MAPE

5.1 Implementation Details

We apply the proposed Modality-Adaptive Pose Estimation(MAPE) method to baseline models to explore its effectiveness in the application of multi-modality HPE. We follow the same training settings of the baseline methods. The initial learning rate is set to 0.001. We decay the learning rate by a factor of 10 at the 170th and 200th epochs. The training concludes at the 210th epoch. We use Adam optimizer for baselines[63, 51, 8, 65] and AdamW for Transformer baselines[67, 66, 70]. All experiments are conducted using PyTorch on NVIDIA GeForce RTX 3090 GPUs.

Table 4. Comparisons between baselines and our proposed method on RAI-MM. The baseline models used are trained on COCO. Models marked with "+" are trained on COCO-IRS. Models marked with "*" and MAPE are trained on COCO-MM.

Method	Backbone	Input size	AP_v	AR_v	AP_{ir}	AR_{ir}	$mmAP$	$mmAR$
SBL [63]	ResNet101	256×192	72.1	74.9	60.6	65.6	66.4	70.3
HRNet [51]	HRNet-W32	256×192	73.3	77.1	60.6	66.4	67.0	71.8
HRNet [51]	HRNet-W48	256×192	75.9	78.9	60.7	66.7	68.3	72.8
HRNet [51]	HRNet-W48	384×288	76.3	79.4	65.8	70.3	71.1	74.9
HrHRNet [8]	HrHRNet-W32	512×512	66.1	69.2	50.2	57.1	58.2	63.2
HRFormer-B [67]	HRFormer-B	384×288	77.0	80.2	66.2	70.8	71.6	75.5
ViTPose-B [66]	ViT-B	256×192	76.2	79.3	65.5	70.0	70.9	74.7
FEPose-B [70]	FEPose-B	256×192	75.8	79.1	66.1	69.7	71.0	74.4
InfPose [65]	InfPose	512×512	68.2	71.3	52.4	57.1	60.3	64.2
SBL+	ResNet101	256×192	61.7	64.7	66.5	70.1	64.1	67.4
HRNet+	HRNet-W32	256×192	64.0	67.4	68.9	72.3	66.5	69.9
HRNet+	HRNet-W48	256×192	65.3	68.7	69.7	72.5	67.5	70.6
HRNet+	HRNet-W48	384×288	66.9	70.1	71.9	75.3	69.4	72.7
HrHRNet+	HrHRNet-W32	512×512	52.8	58.1	58.1	63.5	55.5	60.8
HRFormer-B+	HRFormer-B	256×192	66.5	69.7	71.8	75.2	69.2	72.5
HRFormer-B+	HRFormer-B	384×288	67.5	70.6	72.4	76.3	70.0	73.5
ViTPose-B+	ViT-B	256×192	66.9	70.2	72.0	75.1	69.5	72.7
FEPose-B+	FEPose-B	256×192	66.3	69.5	71.9	75.4	69.1	72.4
InfPose+	InfPose	512×512	53.0	58.9	60.0	64.1	56.5	61.8
SBL*	ResNet101	256×192	68.5	71.9	64.8	68.8	66.7	70.4
HRNet*	HRNet-W32	256×192	69.8	73.9	65.2	70.1	67.5	72.0
HRNet*	HRNet-W48	256×192	71.5	74.9	65.8	71.1	68.7	73.0
HRNet*	HRNet-W48	384×288	72.3	75.8	70.3	74.1	71.3	75.0
HrHRNet*	HrHRNet-W32	512×512	64.5	65.3	55.7	61.3	60.1	63.3
HRFormer-B*	HRFormer-B	256×192	71.9	75.5	70.1	73.5	71.0	74.5
HRFormer-B*	HRFormer-B	384×288	73.0	76.5	71.0	74.8	72.0	75.7
ViTPose-B*	ViT-B	256×192	72.2	75.7	70.1	73.9	71.2	74.8
FEPose-B*	FEPose-B	256×192	71.8	75.6	70.2	73.9	71.0	74.8
InfPose*	InfPose	512×512	65.3	66.8	57.9	63.7	61.6	65.3
MAPE	ResNet101	256×192	75.4	77.3	68.4	71.2	71.9(↑5.5)	74.3
MAPE	HRNet-W32	256×192	76.7	79.4	69.1	72.4	72.9(↑5.9)	75.9
MAPE	HRNet-W48	256×192	77.1	79.8	69.3	72.5	73.2(↑4.9)	76.2
MAPE	HRNet-W48	384×288	77.4	79.9	71.6	74.9	74.5(↑3.4)	77.4
MAPE	HrHRNet-W32	512×512	66.7	71.9	58.7	63.8	62.7(↑4.5)	67.9
MAPE	HRFormer-B	384×288	77.9	81.0	72.1	75.5	75.0(↑3.4)	78.3
MAPE	ViT-B	256×192	77.3	80.0	71.3	74.6	74.3(↑3.4)	77.3
MAPE	FEPose-B	256×192	77.2	79.8	71.2	74.5	74.2(↑3.2)	77.2
MAPE	InfPose	256×192	68.3	72.5	61.2	65.3	64.8(↑4.5)	68.9

5.2 Quantitative Results

To validate the effectiveness of the proposed method on real multi-modality benchmarks, we conduct experiments on the RAI-MM benchmark. The models used are trained on the COCO-MM training set. As shown in Table 4, we find that our method significantly improves the model’s performance in real multi-modality scenarios, highlighting that proposed multimodal datasets have sufficient quality to help the model achieve multimodal HPE capabilities and demonstrating the effectiveness of the proposed method in the real world. It is worth noting that the visible images in RAI-MM include both normal-light scenes and low-light scenes. The results in Table 4 indicate the proposed method can improve the model’s performance under varying lighting conditions to a certain extent.

To further investigate the role of infrared images under different light conditions, we conduct experiments separately on the well-light portion and the low-light portion of the RAI-MM in Table 5. It can be seen that the introduction of infrared modality significantly improves performance in low-light conditions. Moreover, the proposed MAPE effectively mitigates the impact of introducing infrared modality on the performance in well-light scenarios.

To demonstrate the performance improvement on large validation sets, the improvements on the multi-modality benchmarks COCO-MM and MPII-MM are reported in Table 6 and Table 7, respectively. Under the same training dataset, our proposed method significantly enhances the model’s performance on both multi-modality and individual modality benchmarks, while maintaining or improving performance on the individual modality benchmarks compared with models trained on individual modality.

We make further tests on the public infrared dataset UCH [50]. Table 8, where the train set and method used are indicated, shows consistent results with the existing results, confirming the efficacy of our datasets and method in infrared modality HPE.

We compare the proposed MAPE with domain adaption methods [15, 59] and HPE domain adaption methods [60, 29]. We train them on COCO-MM and test them on RAI-MM. For DANN [15] and AdvEnt [59], we assign visible images to a source domain and infrared images to a target domain. AdvMix [60] proposes adversarial training to enhance the robustness of the model and employ knowledge distillation to maintain the performance on clean data. ExlPose [29] proposes adopting learning using privileged information (LUPI) to provide privileged information from visible images to low-light images, enhancing the model’s performance in low-light scenarios. We apply the methods mentioned to the proposed multi-modality benchmark. For AdvMix, we use adversarial training and utilize knowledge distillation from visible images to infrared-style images. For ExlPose, we use LUPI from visible images to infrared-style images. The results are shown in Table 9, which indicate that using one-way information transfer domain adaption is unsatisfactory because visible and infrared images are two modalities that complement each other.

Table 5. Comparisons between baselines and our proposed method on RAI-MM under different light conditions. The baseline models used are trained on COCO. Models marked with "+" are trained on COCO-IRS. Models marked with "*" and MAPE are trained on COCO-MM. AP_{vw} and AP_{vl} represent the performance on visible images in well-light scenes and low-light scenes, respectively.

Method	Backbone	Input size	AP_{vw}	AP_{vl}
SBL [63]	ResNet101	256×192	79.2	55.5
HRNet [51]	HRNet-W32	256×192	80.8	56.3
HrHRNet [8]	HrHRNet-W32	512×512	72.9	50.8
HRFormer-B [67]	HRFormer-B	256×192	83.6	58.1
ViTPose-B [66]	ViT-B	256×192	83.8	58.7
FEPose-B [70]	FEPose-B	256×192	82.8	57.7
InfPose [65]	InfPose	512×512	74.7	52.1
SBL+	ResNet101	256×192	62.6	59.7
HRNet+	HRNet-W32	256×192	65.3	61.8
HrHRNet+	HrHRNet-W32	512×512	53.8	51.1
HRFormer-B+	HRFormer-B	256×192	67.6	63.8
ViTPose-B+	ViT-B	256×192	67.7	64.8
FEPose-B+	FEPose-B	256×192	67.1	64.2
InfPose+	InfPose	512×512	53.9	52.4
SBL*	ResNet101	256×192	70.5	63.9
HRNet*	HRNet-W32	256×192	72.2	65.0
HrHRNet*	HrHRNet-W32	512×512	66.6	60.1
HRFormer-B*	HRFormer-B	256×192	74.1	66.8
ViTPose-B*	ViT-B	256×192	74.2	67.5
FEPose-B*	FEPose-B	256×192	73.7	67.0
InfPose*	InfPose	512×512	66.9	60.6
MAPE	ResNet101	256×192	79.1	66.9
MAPE	HRNet-W32	256×192	80.9	67.8
MAPE	HrHRNet-W32	512×512	71.2	60.4
MAPE	HRFormer-B	256×192	83.2	67.9
MAPE	ViT-B	256×192	83.4	68.6
MAPE	FEPose-B	384×288	81.8	68.4
MAPE	InfPose	256×192	73.4	60.8

Table 6. Comparisons between baselines and our proposed method on COCO-MM. The baseline models used are trained on COCO. Models marked with "*" and MAPE are trained on COCO-MM. For top-down approaches, results are obtained with detected bounding boxes of [51]. Significant improvements are achieved in all metrics.

Method	Backbone	Input size	AP_v	AR_v	AP_{ir}	AR_{ir}	$mmAP$	$mmAR$
SBL [63]	ResNet101	256×192	71.4	77.1	38.8	43.7	55.1	60.4
HRNet [51]	HRNet-W32	256×192	74.4	79.8	44.7	51.9	59.6	65.9
HRNet [51]	HRNet-W48	256×192	75.1	80.4	45.1	52.0	60.1	66.2
HRNet [51]	HRNet-W48	384×288	76.3	81.2	47.2	53.7	61.8	67.5
HrHRNet [8]	HrHRNet-W48	512×512	67.1	72.3	34.2	38.9	50.7	61.5
ViTPose-B [66]	ViT-B	256×192	75.8	81.1	46.5	53.4	61.2	67.3
SBL*	ResNet101	256×192	70.2	75.3	59.4	61.2	64.8	68.3
HRNet*	HRNet-W32	256×192	73.2	76.1	60.5	63.9	66.9	70.0
HRNet*	HRNet-W48	256×192	74.0	76.3	62.9	65.6	68.5	71.0
HRNet*	HRNet-W48	384×288	75.4	77.9	65.8	68.9	70.6	73.4
HrHRNet*	HrHRNet-W32	512×512	65.1	66.9	47.5	56.6	56.3	61.8
ViTPose-B*	ViT-B	256×192	75.1	77.3	65.2	68.1	70.2	72.7
MAPE	ResNet101	256×192	71.3	77.0	61.3	67.4	66.3	72.2
MAPE	HRNet-W32	256×192	73.4	79.0	62.1	68.1	67.8	73.6
MAPE	HRNet-W48	256×192	76.1	78.8	65.0	68.3	70.6	73.6
MAPE	HRNet-W48	384×288	76.5	80.1	66.3	71.9	71.4	76.0
MAPE	HrHRNet-W32	512×512	66.4	69.8	50.9	57.6	58.7	63.7
MAPE	ViT-B	256×192	76.3	79.5	66.0	71.4	71.2	75.5

Table 7. Comparisons between baselines and our method on MPII-MM. The baseline models used are trained on MPII. Models marked with "*" and MAPE are trained on MPII-MM. Significant improvements are obtained in all metrics.

Method	Backbone	Input size	$PCKh_v$	$PCKh_{ir}$	$mmPCKh$
SBL [63]	ResNet101	256×256	89.1	60.1	74.6
HRNet [51]	HRNet-W32	256×256	89.3	67.9	78.6
HRNet [51]	HRNet-W48	256×256	89.6	68.5	79.1
ViTPose-B [66]	ViT-B	256×192	93.3	72.1	82.7
SBL*	ResNet101	256×256	87.0	77.5	82.3
HRNet*	HRNet-W32	256×256	89.4	83.1	86.3
HRNet*	HRNet-W48	256×256	90.0	83.6	86.7
ViTPose-B*	ViT-B	384×288	91.8	84.9	88.4
MAPE	ResNet101	256×256	87.8	81.5	84.7
MAPE	HRNet-W32	256×256	89.9	83.5	86.7
MAPE	HRNet-W48	256×256	90.6	84.3	87.5
MAPE	ViT-B	256×192	93.4	85.6	89.5

Table 8. Performance of HPE methods on UCH.

Metrics	VIS	IRS	MAPE	SBL-Res101	HRNet-W32	HRNet-W48	HrHRNet
AP_{ir}	✓			59.1	60.3	61.2	45.1
AP_{ir}		✓		68.5	69.1	70.4	61.2
AP_{ir}	✓	✓		63.2	64.6	70.0	59.7
AP_{ir}	✓	✓	✓	71.3	72.7	77.0	62.9

Table 9. Comparisons between MAPE and SOTA domain adaption methods on RAI-MM using HRNet-W48 with an input size of 384×288 as the baseline. The models are all trained on COCO-MM.

Method	AP_v	AR_v	AP_{ir}	AR_{ir}	$mmAP$	$mmAR$
Baseline	72.3	75.8	70.3	74.1	71.3	75.0
DANN [15]	75.2	77.8	70.4	74.1	72.8	76.0
AdvEnt [59]	75.7	77.9	70.5	74.2	73.1	76.1
AdvMix [60]	76.1	78.3	70.7	74.4	73.4	76.4
Exlpse [29]	77.0	79.2	71.1	74.5	74.1	76.9
MAPE	77.4	79.9	71.6	74.9	74.5	77.4

5.3 Ablation Studies

Ablation on the network architecture. To verify the effectiveness of different modules of our method, we conduct experiments on the RAI-MM dataset, employing an HRNet-W48 backbone with a size of 384×288 . The methods used are all trained on COCO-MM training set. As shown in Table 10, MSBN and MAL represent variants trained with Modality-Specific Batch Normalization (MSBN) and Modality-Adaptive Loss (MAL), respectively. MAPE signifies the use of our method. We use $diAP = |AP_v - AP_{ir}|$ and $diAR = |AR_v - AR_{ir}|$ to denote the accuracy difference between visible and infrared modalities.

Table 10. Ablation on the network architecture.

Method	AP_v	AR_v	AP_{ir}	AR_{ir}	$mmAP$	$mmAR$	$diAP$	$diAR$
Baseline	72.3	75.8	70.3	74.1	71.3	75.0	2.0	1.7
+MAL	72.0	75.4	70.6	74.2	71.3	74.8	1.4	1.2
+MSBN	77.1	79.6	71.1	74.4	74.1	77.0	6.0	5.2
+MAPE	77.4	79.9	71.6	74.9	74.5	77.4	5.8	5.0

We observe that when using the proposed module individually, MAL reduces the performance gap between individual modalities by 0.6 in $diAP$ and 0.5 in $diAR$, while MSBN shows significant improvement in the multi-modality by 2.8 in $mmAP$ and 2.0 in $mmAR$. Furthermore, we find that utilizing MAL alone fails to achieve satisfactory performance in the visible modality. However, when combined with MSBN, there is a noticeable enhancement by 3.2 in $mmAP$ and 2.4 in $mmAR$. This underscores the importance of MSBN in addressing disparities between visible and infrared modalities. We posit that MSBN effectively prevents modality confusion by learning modality-specific information through distinct batch normalizations. Moreover, it gradually diminishes the modality-specific information during training, effectively bridging the gap between visible and infrared modalities. MAL effectively handles the modality features outputted by MSBN and facilitates the interaction of complementary features between visible and infrared, which accomplishes cross-modality feature transfer and enhances the model’s performance while reducing disparities in modality features.

In short, MSBN effectively handles specific information from each modality, leading to performance improvements across modalities. Meanwhile, MAL facilitates effective interaction between modalities, reducing the inter-modality differences by MSBN processing. The combination of MSBN and MAL achieves effective modality adaptation.

Ablation on the direction of modality interaction in MAL. The methods used are trained on COCO-MM and tested on RAI-MM. The results shown in Table 11 suggest that one-way interaction between visible and infrared modalities fails to produce satisfactory results. This indicates that the two modalities can complement each other and proves the rationality of MAL, which operates on both modalities.

Table 11. Ablation on direction of modality interaction in MAL.

Method	AP_v	AR_v	AP_{ir}	AR_{ir}	$mmAP$	$mmAR$
Baseline	72.3	75.8	70.3	74.1	71.3	75.0
$vis \rightarrow ir$	74.8	77.3	67.6	70.4	71.2	73.9
$ir \rightarrow vis$	71.2	74.4	71.3	74.5	71.3	74.5
$vis \leftrightarrow ir(ours)$	77.4	79.9	71.6	74.9	74.5	77.4

5.4 Qualitative Comparison

Here we provide qualitative comparisons between our proposed method and the baseline approach on proposed benchmarks. Figure 10 shows comparisons on COCO-MM, where we use HRNet-W48 with an input size of 384×288 as the baseline. Figure 11 shows comparisons on MPII-MM, where we use HRNet-W32 with an input size of 256×256 as the baseline. Results show that while the baseline method performs well in estimating poses within the visible modality, its performance significantly deteriorates in the infrared modality. Although training the baseline on multimodal data moderately enhances its performance in the infrared modality, it still falls short of achieving precise pose estimations. In contrast, our proposed method consistently achieves superior pose estimation results in both the visible and infrared modalities.

Figure 12 shows the comparisons on RAI-MM, where we use HRNet-W48 with an input size of 384×288 as the baseline. It is worth noting that the models used are trained on COCO-MM and RAI-MM is only used for testing. It can be observed that our proposed dataset and method consistently improve the performance of HPE in real infrared images. Notably, the visible images in RAI-MM include low-light scenes, and after incorporating infrared images and proposed method, the performance in low-light conditions also improves. Such improvement is crucial for practical applications of HPE under different light conditions.

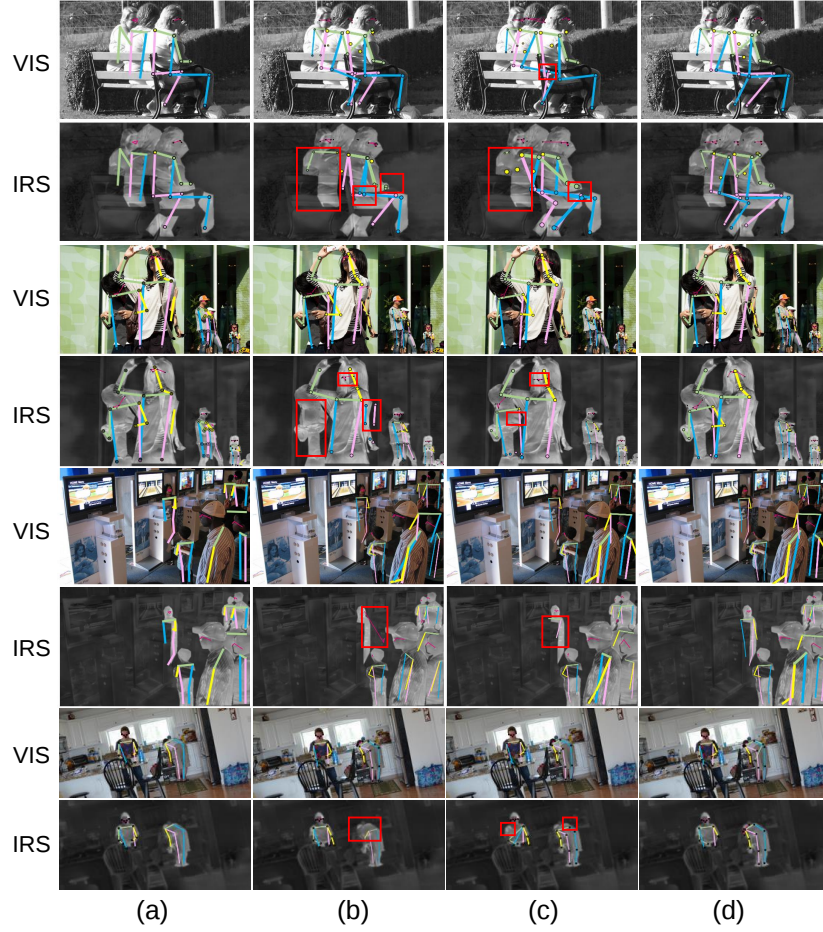


Fig. 10. Qualitative comparisons on the COCO-MM dataset. The predicted poses in the visible(VIS) and infrared-style(IRS) modalities are displayed on the corresponding images. (a)Ground truth. (b)Predictions from baseline trained on COCO. (c)Predictions from baseline trained on COCO-MM. (d)Predictions from proposed MAPE trained on COCO-MM. Red boxes indicate the areas where keypoint detection errors occur.

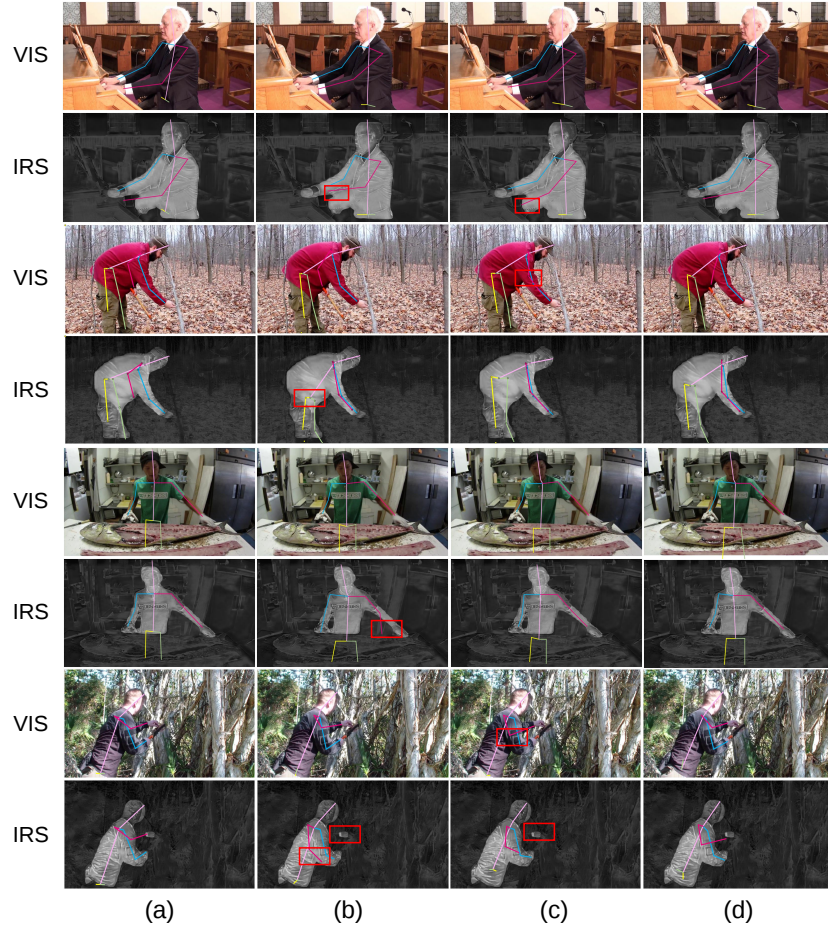


Fig. 11. Qualitative comparisons on the MPII-MM dataset. The predicted poses in the visible(VIS) and infrared-style(IRS) modalities are displayed on the corresponding images. (a)Ground truth. (b)Predictions from baseline trained on MPII. (c)Predictions from baseline trained on MPII-MM. (d)Predictions from proposed MAPE trained on MPII-MM. Red boxes indicate the areas where keypoint detection errors occur.

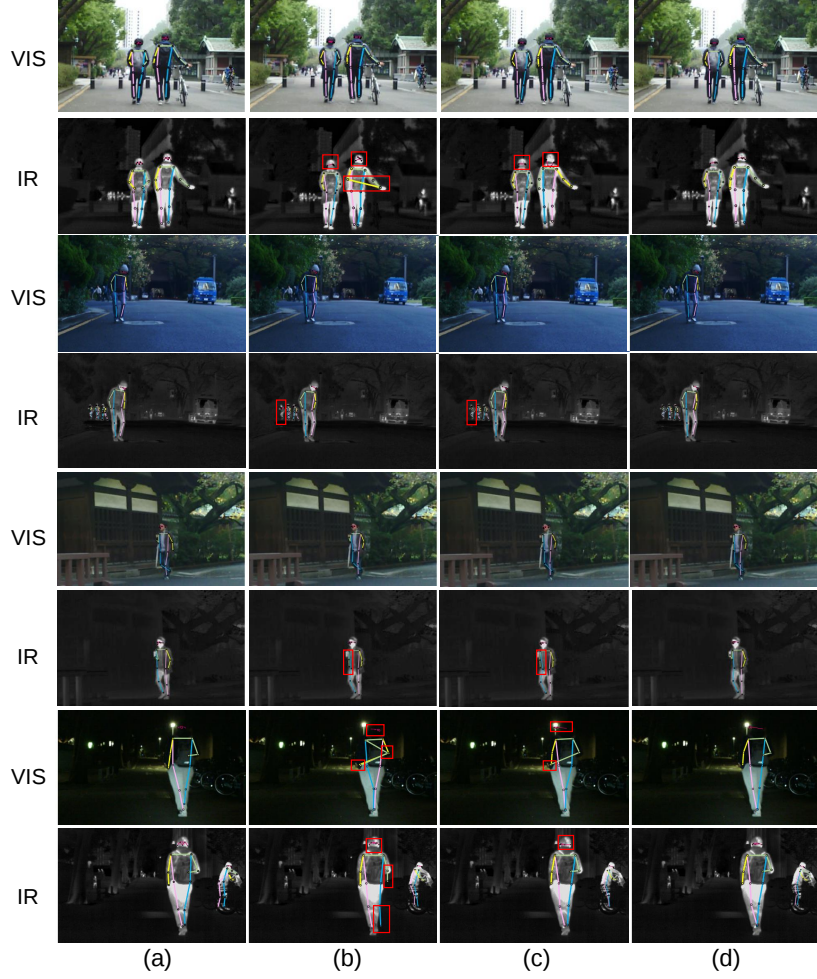


Fig. 12. Qualitative comparisons on the RAI-MM dataset. The predicted poses in the visible(VIS) and infrared(IR) modalities are displayed on the corresponding images. (a)Ground truth. (b)Predictions from baseline trained on COCO. (c)Predictions from baseline trained on COCO-MM. (d)Predictions from proposed MAPE trained on COCO-MM. Red boxes indicate the areas where keypoint detection errors occur.

6 Conclusion

In summary, this paper addresses the inherent limitations of single modality human pose estimation (HPE) approaches by introducing innovative techniques and benchmark datasets to advance multimodal HPE. Our proposed Instance Cross-modality Style Transfer Network(ICSTN) facilitates the conversion of visible images to infrared style images, enhancing the realism of infrared HPE, based on which we introduce a novel visible-infrared benchmark comprising two large-scale generated datasets COCO-MM, MPII-MM, and one real-world dataset RAI-MM for assessing multimodal HPE. By testing the existing methods on multimodal setting, we find that performance degrades severely. We then propose Modality-Adaptive Pose Estimation (MAPE) demonstrating superior performance across both visual and infrared modalities. By seamlessly integrating into existing methodologies and addressing challenges posed by modality variance, MAPE sets a new baseline for multimodal HPE.

Acknowledgment. This research was supported by Ningbo 2025 Science and Technology Innovation Major Project (No. 2022Z072, No.2023Z044).

References

1. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5167–5176 (2018)
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3686–3693 (2014)
3. Arar, M., Ginger, Y., Danon, D., Bermano, A.H., Cohen-Or, D.: Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13410–13419 (2020)
4. Artacho, B., Savakis, A.: Unipose: Unified human pose estimation in single images and videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7035–7044 (2020)
5. Cao, Y., Bin, J., Hamari, J., Blasch, E., Liu, Z.: Multimodal object detection by channel switching and spatial attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 403–411 (2023)
6. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4733–4742 (2016)
7. Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 7354–7362 (2019)
8. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5386–5395 (2020)

9. Cormier, M., Clepe, A., Specker, A., Beyerer, J.: Where are we with human pose estimation in real-world surveillance? In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 591–601 (2022)
10. Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.D., et al.: A survey on multimodal large language models for autonomous driving. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 958–979 (2024)
11. Desmarais, Y., Mottet, D., Slangen, P., Montesinos, P.: A review of 3d human pose estimation algorithms for markerless motion capture. *Computer Vision and Image Understanding* **212**, 103275 (2021)
12. Di, W., Jinyuan, L., Xin, F., Liu, R.: Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In: *International Joint Conference on Artificial Intelligence (IJCAI)* (2022)
13. Dubey, S., Dixit, M.: A comprehensive survey on human pose estimation approaches. *Multimedia Systems* **29**(1), 167–195 (2023)
14. Feng, R., Gao, Y., Tse, T.H.E., Ma, X., Chang, H.J.: Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 14861–14872 (2023)
15. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.S.: Domain-adversarial training of neural networks. In: *Journal of machine learning research* (2015), <https://api.semanticscholar.org/CorpusID:2871880>
16. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2414–2423 (2016)
17. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14676–14686 (2021)
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2961–2969 (2017)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
21. Holmquist, K., Wandt, B.: Diffpose: Multi-hypothesis human pose estimation using diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 15977–15987 (2023)
22. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International conference on machine learning*. pp. 448–456 (2015)
23. Jiang, K., Zhang, T., Liu, X., Qian, B., Zhang, Y., Wu, F.: Cross-modality transformer for visible-infrared person re-identification. In: *European Conference on Computer Vision*. pp. 480–496. Springer (2022)
24. Jin, S., Liu, W., Xie, E., Wang, W., Qian, C., Ouyang, W., Luo, P.: Differentiable hierarchical graph grouping for multi-person pose estimation. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII* 16. pp. 718–734. Springer (2020)
25. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: *British Machine Vision Conference*. p. 5 (2010)

26. Kocabas, M., Karagoz, S., Akbas, E.: Multiposenet: Fast multi-person pose estimation using pose residual network. In: Proceedings of the European conference on computer vision (ECCV). pp. 417–433 (2018)
27. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11977–11986 (2019)
28. Lee, S.P., Kini, N.P., Peng, W.H., Ma, C.W., Hwang, J.N.: Hupr: A benchmark for human pose estimation using millimeter wave radar. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5715–5724 (2023)
29. Lee, S., Rim, J., Jeong, B., Kim, G., Woo, B., Lee, H., Cho, S., Kwak, S.: Human pose estimation in extremely low-light conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 704–714 (2023)
30. Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., Tu, Z.: Pose recognition with cascade transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1944–1953 (2021)
31. Li, Y., Yang, S., Liu, P., Zhang, S., Wang, Y., Wang, Z., Yang, W., Xia, S.T.: Simcc: A simple coordinate classification perspective for human pose estimation. In: European Conference on Computer Vision. pp. 89–106. Springer (2022)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
33. Lin, W., Liu, H., Liu, S., Li, Y., Xiong, H., Qi, G., Sebe, N.: Hieve: A large-scale benchmark for human-centric video analysis in complex events. *International Journal of Computer Vision* **131**(11), 2994–3018 (2023)
34. Liu, H., Chen, Y., Zhao, W., Zhang, S., Zhang, Z.: Human pose recognition via adaptive distribution encoding for action perception in the self-regulated learning process. *Infrared Physics & Technology* **114**, 103660 (2021)
35. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5802–5811 (2022)
36. Liu, S., Huang, X., Fu, N., Li, C., Su, Z., Ostadabbas, S.: Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(1), 1106–1118 (2022)
37. Liu, Z., Feng, R., Chen, H., Wu, S., Gao, Y., Gao, Y., Wang, X.: Temporal feature alignment and mutual information maximization for video-based human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11006–11016 (2022)
38. Luvizon, D.C., Picard, D., Tabia, H.: 2d/3d pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5137–5146 (2018)
39. Luvizon, D.C., Tabia, H., Picard, D.: Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics* **85**, 15–22 (2019)
40. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of machine learning research* **9**(11) (2008)

41. Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z., den Hengel, A.v.: Poseur: Direct human pose regression with transformers. In: European Conference on Computer Vision. pp. 72–88. Springer (2022)
42. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging* **34**(10), 1993–2024 (2014)
43. Miller, J.: Principles of Infrared Technology: A Practical Guide to the State of the Art. Springer US (2012), <https://books.google.com/books?id=b2vVBwAAQBAJ>
44. Mollyn, V., Arakawa, R., Goel, M., Harrison, C., Ahuja, K.: Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2023)
45. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems* **30** (2017)
46. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14. pp. 483–499. Springer (2016)
47. Qu, H., Cai, Y., Foo, L.G., Kumar, A., Liu, J.: A characteristic function-based method for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13009–13018 (2023)
48. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18. pp. 234–241 (2015)
49. Sapp, B., Taskar, B.: Modex: Multimodal decomposable models for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3674–3681 (2013)
50. Smith, J., Loncomilla, P., del Solar, J.R.: Human pose estimation using thermal images. *IEEE Access* (2023)
51. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019)
52. Sun, P., Gu, K., Wang, Y., Yang, L., Yao, A.: Rethinking visibility in human pose estimation: Occluded pose reasoning via transformers. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5903–5912 (2024)
53. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European conference on computer vision (ECCV). pp. 529–545 (2018)
54. Sun, Y., Dougherty, A.W., Zhang, Z., Choi, Y.K., Wu, C.: Mixsynthformer: A transformer encoder-like structure with mixed synthetic self-attention for efficient human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14884–14893 (2023)
55. Tang, L., Deng, Y., Ma, Y., Huang, J., Ma, J.: Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica* **9**(12), 2121–2137 (2022)
56. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1653–1660 (2014)

57. Vachmanus, S., Ravankar, A.A., Emaru, T., Kobayashi, Y.: Multi-modal sensor fusion-based semantic segmentation for snow driving scenarios. *IEEE sensors journal* **21**(15), 16839–16851 (2021)
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
59. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation (2019), <https://arxiv.org/abs/1811.12833>
60. Wang, J., Jin, S., Liu, W., Liu, W., Qian, C., Luo, P.: When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 11855–11864 (2021)
61. Wang, Y., Li, M., Cai, H., Chen, W.M., Han, S.: Lite pose: Efficient architecture design for 2d human pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13126–13136 (2022)
62. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 466–481 (2018)
63. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 466–481 (2018)
64. Xu, H., Ma, J., Le, Z., Jiang, J., Guo, X.: FusionDn: A unified densely connected network for image fusion. In: *Proceedings of the AAAI conference on artificial intelligence*. pp. 12484–12491 (2020)
65. Xu, X., Wei, X., Xu, Y., Zhang, Z., Gong, K., Li, H., Xiao, L.: Infpose: Real-time infrared multi-human pose estimation for edge devices based on encoder-decoder cnn architecture. *IEEE Robotics and Automation Letters* (2023)
66. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation (2022), <https://arxiv.org/abs/2204.12484>
67. Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: Hrformer: High-resolution transformer for dense prediction (2021), <https://arxiv.org/abs/2110.09408>
68. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 7093–7102 (2020)
69. Zheng, J., Shi, X., Gorban, A., Mao, J., Song, Y., Qi, C.R., Liu, T., Chari, V., Cornman, A., Zhou, Y., et al.: Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4478–4487 (2022)
70. Zhu, Z., Dong, W., Gao, X., Peng, A.: Towards infrared human pose estimation via transformer. In: *2023 International Joint Conference on Neural Networks (IJCNN)*. pp. 1–8 (2023). <https://doi.org/10.1109/IJCNN54540.2023.10191173>