electronics

MDPI

Article

NRPerson: A Non-Registered Multi-Modal Benchmark for Tiny Person Detection and Localization

Yi Yang †**, Xumeng Han** † **, Kuiran Wang , Xuehui Yu, Wenwen Yu, Zipeng Wang, Guorong Li , Zhenjun Han * and Jianbin Jiao**

School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China; yysolon@163.com (Y.Y.); hanxumeng19@mails.ucas.ac.cn (X.H.); wangkuiran19@mails.ucas.ac.cn (K.W.); yuxuehui17@mails.ucas.ac.cn (X.Y.); yuwenwen22@mails.ucas.ac.cn (W.Y.); wangzipeng22@mails.ucas.ac.cn (Z.W.); liguorong@ucas.ac.cn (G.L.); jiaojb@ucas.ac.cn (J.J.)

* Correspondence: hanzhj@ucas.ac.cn

† These authors contributed equally to this work.

Abstract: In recent years, the detection and localization of tiny persons have garnered significant attention due to their critical applications in various surveillance and security scenarios. Traditional multi-modal methods predominantly rely on well-registered image pairs, necessitating the use of sophisticated sensors and extensive manual effort for registration, which restricts their practical utility in dynamic, real-world environments. Addressing this gap, this paper introduces a novel non-registered multi-modal benchmark named NRPerson, specifically designed to advance the field of tiny person detection and localization by accommodating the complexities of real-world scenarios. The NRPerson dataset comprises 8548 RGB-IR image pairs, meticulously collected and filtered from 22 video sequences, enriched with 889,207 high-quality annotations that have been manually verified for accuracy. Utilizing NRPerson, we evaluate several leading detection and localization models across both mono-modal and non-registered multi-modal frameworks. Furthermore, we develop a comprehensive set of natural multi-modal baselines for the innovative non-registered track, aiming to enhance the detection and localization of unregistered multi-modal data using a cohesive and generalized approach. This benchmark is poised to facilitate significant strides in the practical deployment of detection and localization technologies by mitigating the reliance on stringent registration requirements.

Keywords: non-registered; multi-modal; tiny person detection/localization

check for updates

Citation: Yang, Y.; Han, X.; Wang, K.; Yu, X.; Yu, W.; Wang, Z.; Li, G.; Han, Z.; Jiao, J. NRPerson: A Non-Registered Multi-Modal Benchmark for Tiny Person Detection and Localization. *Electronics* **2024**, *13*, 1697. https://doi.org/10.3390/electronics13091697

Academic Editor: Stefanos Kollias

Received: 30 March 2024
Revised: 17 April 2024
Accepted: 22 April 2024
Published: 27 April 2024

1. Introduction

Person detection and localization is a critical research topic in computer vision, fundamental to numerous higher-level applications such as surveillance [1–4], tracking [5–8], and rapid rescue operations [9,10]. Despite longitudinal studies and significant progress, person detection continues to face challenges such as small object scale, variable lighting, and complex backgrounds, which severely hinder the accuracy and reliability of detection systems.

Detecting and localizing tiny-sized objects within an image has become a focus for many computer vision researchers due to the substantial technical challenges and significant application value it presents [9–15]. The detection of tiny persons is especially crucial in wide-area and long-range scenarios, where precise identification can aid in critical tasks such as monitoring crowded events or coordinating rescue missions in expansive environments.

Current studies predominantly utilize mono-modal RGB images for detection tasks [9–20]. However, this approach often fails to address all environmental challenges, as RGB data alone may not provide sufficient information under poor lighting or when objects blend into complex backgrounds.

Multi-modal data, integrating inputs from various sensor types such as RGB and infrared, have proven effective in enhancing mono-modal detection tasks [3,21–29]. A common practice in multi-modal person detection tasks assumes that the image pairs are well registered [21,30–35]. However, this requirement for precise sensor alignment is often impractical in dynamic real-world conditions and leads to significant challenges in deployment due to the complex registration processes required.

Recognizing these limitations, we introduce NRPerson, a novel benchmark for non-registered multi-modal tiny person detection and localization. This benchmark aims to explore the utility of multi-modal image pairs without the stringent necessity for exact registration, thus more closely mimicking real-world scenarios, where perfect alignment of sensors cannot be guaranteed. NRPerson comprises 8548 RGB-IR image pairs collected from 22 video sequences, supported by 889,207 high-quality annotations that are manually verified to ensure their accuracy and reliability (see Figure 1).

This approach not only simplifies the data collection process by reducing dependency on sophisticated registration techniques but also enhances the practical application of detection systems in diverse operational environments. We evaluate several leading detection models on both mono-modal and non-registered multi-modal tracks to assess their effectiveness in leveraging non-registered data for improved performance.

The contributions of this paper are significant in advancing the field of multi-modal person detection:

1. We establish a large-scale NRPerson benchmark, introducing the concept of non-registration in multi-modal data handling for tiny person detection and localization, thereby promoting tasks that align closer to real-world scenarios.
2. We perform a comprehensive evaluation of state-of-the-art detection and localization models on this new benchmark, demonstrating the potential and limitations of existing technologies when applied to non-registered multi-modal data.
3. We develop a set of diverse natural multi-modal baselines that effectively utilize non-registered data, laying a robust foundation for future research in multi-modal detection and localization.

In the subsequent sections, we detail the methodology for constructing the NRPerson dataset, describe our evaluation framework for detection models, and discuss the broader implications of our findings for the field.

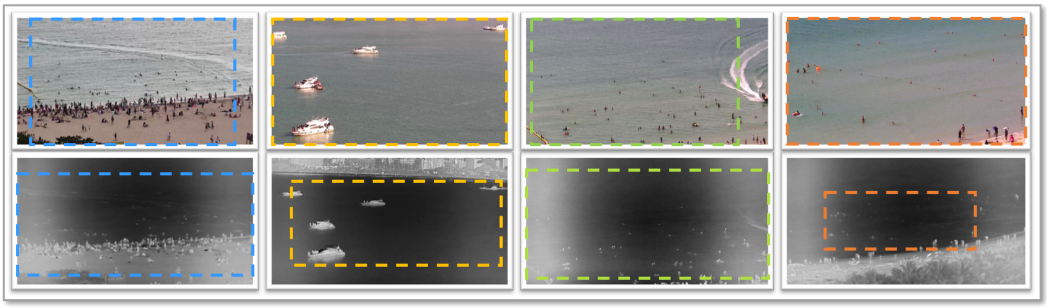

Figure 1. Examples of NRPerson. They are maritime aerial scenes containing a large number of tiny persons. Each column represents a time-aligned RGB-IR image pair. The dashed box on each image pair represents a consistent field of view, showing that the image pairs are spatial-overlapped and non-registered.

2. Related Work

2.1. Multi-Modal Person Detection

Multi-modal person detection has become a popular topic in the research community, and many state-of-the-art multi-modal person detectors [21,30–36] have been proposed. ProbEn [37] fuses outputs from multimodal detectors such as RGB and thermal cameras. This approach leverages Bayes' rule to calculate a combined score that more accurately reflects the likelihood of object presence, thereby improving detection performance even in poor illumination—a common challenge in person detection scenarios.

A common assumption for these multi-modal detectors is that the image pairs are well registered. However, in real-world scenarios, the multi-modal registration assumption is hard to come into existence, and research [36] has revealed that even manually calibrated multi-modal data still suffer from weak alignment. Weak alignment is also a relatively ideal situation that still requires complex operation, while the actual data collection situation is more complicated and changeable. Therefore, we establish a non-registered multi-modal benchmark to change the prerequisites for good alignment, thus getting rid of the reliance on manual calibration to some extent.

2.2. Tiny Object Detection

Recently, many researchers have devoted themselves to the study of tiny object detection [9,13,16–20], but research in the field remains limited. The low resolution of tiny objects makes feature representation insufficient, leading to poor performance compared to large objects. CEFP2N [38] introduce a feature pyramid network enhanced by a Context Enhancement Module (CEM) and a Feature Purification Module (FPM) for tiny object detection. Their approach augments context information and purifies features, significantly improving detection accuracy with a novel "copy-reduce-paste" data augmentation method. Accordingly, we are motivated to propose NRPerson, aiming at detection task in multi-modal and long-distance scenarios for tiny persons. Based on NRPerson, multi-modal data can be used to complement instance feature information to alleviate the lack of feature representation capability.

2.3. Point-Based Object Localization

Point-based object localization aims to train with point-level annotations, thereby learning the capability to predict the location coordinates of objects [10,12,39]. Unlike object detection, object localization is not interested in the object's scale. Therefore, costly and time-consuming box-level annotations are unnecessary, and simple and time-saving point-level annotations are more promising and valuable. CPR [10] introduces a general object localization method based on coarse point supervision and provides pseudo boxes generated by refined point annotations. As a result, the point-to-point learning process can be converted to a box-to-box one. CPR++ [40] advances the field of point-based object localization by refining coarse point annotations through a dynamic, multi-stage process that progressively reduces semantic variance, improving localization accuracy. Point-based object localization is in the ascendant, and the proposed NRPerson provides a large-scale benchmark and extends to multi-modal data.

2.4. Other Public Datasets

Representative public datasets that are relevant to NRPerson are summarized in Table 1. They can be divided into three categories: (1) multi-modal datasets [6,22–25,41,42], (2) person detection/localization datasets [1,2,5,9,10,43–45] and (3) multi-modal person detection/localization datasets [3,8,21,46]. OSU [41] is a thermal-visible video fusion dataset for moving target tracking and pedestrian motion analysis and classification. LITIV [6] consists of videos for tracking scenarios captured through thermal and visual cameras with different zoom settings and positions. CityPersons [2] is a rich and diverse pedestrian detection dataset. It is recorded by a car traversing various cities and contains dense pedestrians. TinyPerson [9] is the first benchmark for long-distance tiny person detection. KAIST [21] provides well-aligned color-thermal image pairs collected in various traffic scenes day and night. LLVIP [3] is a visible-infrared paired dataset containing a large number of pedestrians with low brightness. NRPerson is a large-scale multi-modal tiny person detection and localization dataset. Unlike other datasets based on manual registration and provide modality-shared annotations, NRPerson first introduces the concept of non-registration, which is closer to real-world scenarios.

Table 1. Comparisons of the scale and properties of NRPerson with several related datasets. The description of the properties is presented in Section 3.2 in which *non-registration* refers to the data property and annotations based on non-registration image pairs rather than shared across modalities. TA denotes time-aligned, SO denotes spatial-overlapped, and NR denotes non-registered. PD is person detection, PL is person localization, OD is object detection, and OT is object tracking. –: the property is not applicable to this dataset. †: the visible images of CVC-14 are in grayscale. *: some image pairs are not strictly time-aligned.

Dataset	Images	Annotations	RGB	IR	TA	SO	NR	Task	Year
Multi-modal dataset									
OSU-CT [41]	17,088	–	✓	✓	✓	✓	×	Fusion	2007
LITIV [6]	12,650	–	✓	✓	✓	✓	×	Fusion	2012
TNO [42]	480	–	✓	✓	✓	✓	×	Fusion	2014
RGB-T210 [22]	210,000	210,000	✓	✓	✓	✓	×	OT	2017
RGB-T234 [23]	233,800	233,800	✓	✓	✓	✓	×	OT	2019
RegDB [24]	8240	8240	✓	✓	–	–	–	Re-ID	2017
SYSU-MM01 [25]	45,863	45,863	✓	✓	–	–	–	Re-ID	2020
Person detection/localization dataset									
Caltech [1]	249,884	346,621	✓	×	–	–	–	PD	2009
KITTI [5]	14,999	80,256	✓	×	–	–	–	OD	2012
COCOPersons [45]	64,115	273,469	✓	×	–	–	–	PD	2014
CityPersons [2]	5000	35,016	✓	×	–	–	–	PD	2017
CrowdHuman [43]	24,370	456,098	✓	×	–	–	–	PD	2018
SCUT-FIR [44]	211,011	477,907	×	✓	–	–	–	PD	2019
TinyPerson [9]	1610	72,651	✓	×	–	–	–	PD	2019
SeaPerson [10]	12,032	619,627	✓	×	–	–	–	PL	2022
Multi-modal person detection/localization dataset									
KAIST [21]	190,656	103,128	✓	✓	✓	✓	×	PD	2015
CVC-14 [46]	17,002	17,929	✓†	✓	✓*	✓	×	PD	2016
FLIR [8]	28,267	119,491	✓	✓	✓*	✓	×	OD	2018
LLVIP [3]	30,976	41,579	✓	✓	✓	✓	×	PD	2021
NRPerson	17,096	889,207	✓	✓	✓	✓	✓	PD & PL	2022

3. The NRPerson Dataset

3.1. Data Collection and Annotation

Data Collection. The equipment is a binocular camera platform consisting of an RGB camera and an IR camera. As we are committed to exploring the detection challenges in wild maritime quick rescue, images containing many person objects are captured from various seaside scenes. We sample images from 22 video sequences, and discard images with no objects or high homogeneity. After time alignment and manual filtering, time-synchronized and high-quality image pairs containing persons are selected. The resolutions of RGB and IR images are 1920 × 1080 and 960 × 576 pixels, respectively.

Data Annotation. We use three types of labels to annotate objects *person*, *ignore*, and *uncertain*. Distinguishable persons are annotated as *person* by accurate bounding boxes. Areas with dense crowds, reflections, or ambiguities are annotated as *ignore*. Some objects are difficult to recognize as persons, and thus we annotate them as *uncertain*. Among the filtered 8548 RGB-IR image pairs, a total of 889,207 persons with bounding boxes are manually annotated and carefully double-checked, including 471,924 persons in RGB images and 417,283 persons in IR images as shown in Tables 1 and 2. In addition, we follow [10] to generate coarse point annotation in each labeled bounding box.

Table 2. Detailed statistical information for NRPerson.

NRPerson	Train	Valid	Test	Total
image pairs	4614	375	3559	8548
annos	346,413	33,385	509,408	889,207
RGB annos	174,312	16,943	280,669	471,924
IR annos	172,101	16,442	228,739	417,282

Training and Test Sets. The training, validation and test sets are constructed by randomly dividing the images into three subsets, and images from the same video sequence cannot be divided into different subsets. The results of the dataset division are summarized in Table 2. Specifically, the training set contains 8 video sequences with 4614 image pairs and 346,413 annotations, the validation set contains 1 video sequence with 375 image pairs and 33,385 annotations, and the test set contains 13 video sequences with 3559 image pairs and 509,409 annotations.

3.2. Dataset Properties

Time Alignment. RGB and IR video sequences are captured simultaneously using a binocular camera. The two video sequences are temporally aligned through manual calibration. Then, we sample the frames and ensure that the capturing time for each image pair is identical to guarantee strict time alignment.

Spatial Overlapping. Although time-aligned image pairs are not geometrically aligned due to the different physical properties of RGB and IR sensors, the shooting angles of the binocular sensors are very close. Therefore, the image pairs have spatial overlap. Manual filtering is adopted to remove anomalies and ensure that each image pair has spatial overlap.

Non-registration. Non-registration means that RGB-IR image pairs are time-aligned and spatial-overlapped but are not geometrically aligned. Non-registered image pairs have the following three manifestations:

(1) Position shift: Sensor differences and external factors lead to position shifts between corresponding instances in image pairs. This is the most common phenomenon, and even manually registered data often cannot completely eliminate it.

(2) Size difference: Differences in the field of view and image pixels captured by RGB and IR sensors lead to the size difference of the corresponding instances. This can basically be eliminated through manual calibration.

(3) No correspondence: There is no correspondence between the pixel coordinates of non-registered image pairs. Hence, instances between different modalities cannot be matched and paired.

Tiny. Absolute size is defined as the number of pixels in the bounding box area, and relative size is the absolute size divided by the image pixels. As shown in Table 3 and Figure 2, the absolute size and relative size of persons in NRPerson are very small compared to other representative datasets [2,3,8–10,21]. Objects with small absolute scales are blurred and contain less semantic information, which increases the difficulty of the detector to identify the objects and greatly affects the performance [9]. Tiny relative scales result in the scenes having a large number of complex backgrounds, which exacerbates the issues of positive–negative imbalance and false positives, hence being more challenging and exploratory.

Aspect Ratio Diversity. The aspect ratio of persons in NRPerson has a large variance as shown in Table 3. It suggests that unlike conventional person detection/localization datasets with relatively single perspectives and poses, our dataset is an effective complement in terms of aspect ratio diversity, bringing more complexity and making detection and localization more challenging.

Table 3. Mean and standard deviation statistics of absolute size, relative size and aspect ratio of *person* annotations in NRPerson and several representative datasets.

Dataset	Absolute Size	Relative Size	Aspect Ratio
KAIST [21]	54.2 ± 24.8	0.10 ± 0.05	0.45 ± 0.10
FLIR [8]	26.6 ± 23.1	0.05 ± 0.04	0.43 ± 0.19
LLVIP [3]	146.1 ± 32.4	0.13 ± 0.03	0.47 ± 0.19
CityPersons [2]	79.8 ± 67.5	0.56 ± 0.05	0.41 ± 0.01
TinyPerson [9]	18.0 ± 17.4	0.01 ± 0.01	0.68 ± 0.42
SeaPerson [10]	22.6 ± 10.8	0.02 ± 0.01	0.72 ± 0.42
NRPerson	23.7 ± 9.6	0.02 ± 0.01	0.81 ± 0.45

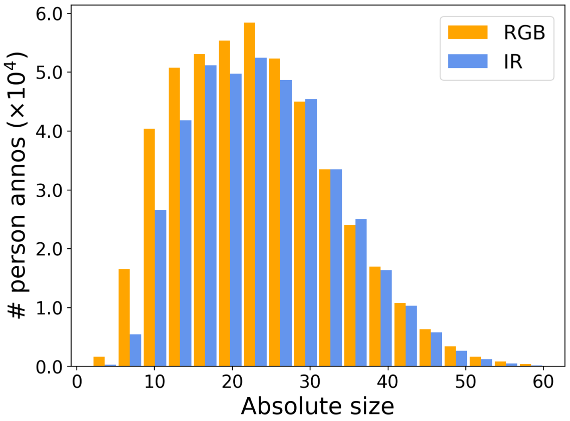

Figure 2. The histogram of the absolute size of person annotations in NRPerson. The statistical results reflect that the object size is concentrated in areas with small values, showing the tiny property of our dataset.

Privacy. Personal privacy is a sensitive matter and an important issue in person-centered research, such as person detection, face detection, and person re-identification. For the NRPerson benchmark, objects are tiny persons with low resolution at long distances (about 24 pixels per person on average, as shown in Table 3). Therefore, little personal information emerges through the data. Additionally, we will enforce dataset terms before releasing NRPerson:

(1) Sharing dataset images on the Internet in any form without permission is prohibited;
(2) Identifiable images of people cannot be used to make demos or promotions;
(3) Use face obfuscation or shield face regions when applicable and study its impact on detection and localization [47].

4. Tracks and Metrics

4.1. Mono-Modal Track

The NRPerson dataset contains many tiny-sized objects, as shown in Figure 1, making it a favorable complement to the existing tiny object detection datasets. In addition, other datasets are basically based on RGB images but rarely on IR images. NRPerson bridges this gap and facilitates the development of research on IR tiny object detection and localization in the community.

We establish a mono-modal track on NRPerson that follows the task settings of generic object detection [48,49] and localization [10,12]. Object detection has been widely studied and is well known, so we do not describe it further. Object localization refers to training the model with point-level annotations to predict the objects' locations. There are various settings of point-level annotations in previous studies [10,12,50]. In this paper, we choose a more generalized annotation, i.e., a single point annotation with a random position on each object.

In addition to benchmarking the performance of detectors and localizers under their respective tasks, we also try to explore the performance of the same model structure under multiple tasks. However, detectors and localizers have different learning supervision and inference results, so it is challenging to perform multi-task unification directly. According to [10], we convert the point-to-point problem into a box-to-box, enabling the detectors to achieve object localization under point supervision. Specifically, CPR [10] is adopted to refine the initial coarse point annotations, and then pseudo boxes are generated with the points as the center that serve as supervision for model training. During the inference, the center points of the predicted boxes are regarded as the localization results. In this way, we build a multi-task mono-modal track on NRPerson to report the detection and localization performance of representative models.

Detection metrics. We adopt average precision (AP) as the evaluation metric for object detection. Since many applications of tiny person detection (e.g., shipwreck search and rescue) focus more on finding the approximate location of a person rather than the exact bounding box coordinates, we adopt AP_{50} as the evaluation metric, i.e., the Intersection over Union (IoU) threshold is set to 0.5. For more detailed experimental comparisons, bounding box sizes are divided into three intervals: $t(tiny) \in [2, 20]$, $s(small) \in (20, 32]$, and $r(reasonable) \in (32, \infty)$.

Localization metrics. We follow the evaluation metric of object localization adopted in [10]. Specifically, the distance d between point $p = (x, y)$ and bounding box $b = (x_c, y_c, w, h)$ is calculated as:

$$d(p, b) = \sqrt{\left(\frac{x - x_c}{w}\right)^2 + \left(\frac{y - y_c}{h}\right)^2}, \tag{1}$$

where (x_c, y_c), w, h are the center point coordinates, width, and height of the bounding box. A point matches a ground-truth bounding box if the distance d is smaller than the threshold τ. If there are multiple matching points through a ground-truth, the point with the highest score is selected; if a point has multiple matching ground-truths, the ground-truth with the smallest distance d is selected. A point is counted as a true positive (TP) when it matches the ground-truth, and a false positive (FP) when it does not. If a point matches an object annotated as *ignore*, neither TP nor FP are counted, following [2,9]. Likewise, we adopt AP_τ as the evaluation metric and set $\tau = 1.0$ as the default.

4.2. Non-Registered Multi-Modal Track

We introduce the concept of non-registration on the basis of multi-modal tiny person detection and localization, thereby establishing a non-registered multi-modal track. The goal is to directly learn from non-registered data such that the model achieves robust detection and localization capabilities in each modality. A flowchart of non-registered multi-modal model is shown in Figure 3.

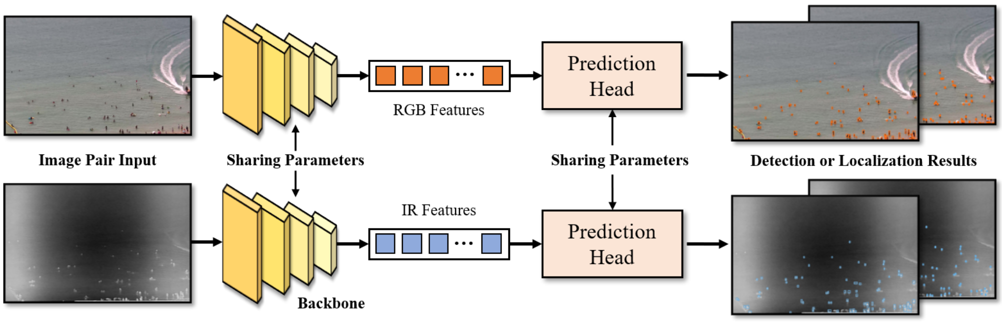

Figure 3. A flowchart of multi-modal model on the non-registered multi-modal track. Multi-modal image pairs are simultaneously input into the model to obtain detection or localization results for each modality. Orange represents RGB features and Blue represents the IR features.

Task Definition. The RGB image set is denoted as $\mathcal{X}_{rgb} = \{x_r^1, x_r^2, \ldots, x_r^n\}$ and the IR image set is denoted as $\mathcal{X}_{ir} = \{x_i^1, x_i^2, \ldots, x_i^n\}$, where x_r^k and x_i^k represent a non-registered RGB-IR image pair satisfying both time-alignment and spatial-overlapping. We require the model $\mathcal{M}$ to input non-registered images in pairs and output the results for each modality, i.e., $(y_r^k, y_i^k) = \mathcal{M}(x_r^k, x_i^k)$, where y_r^k and y_i^k denote the detection/localization results of RGB and IR images, respectively. During the evaluation, the multi-modal test set is used to comprehensively evaluate the performance across all modalities.

Dual-Stream Input. The conventional detectors and localizers are based on a single image input. To meet the requirements of the non-registered multi-modal track, we adopt a dual-stream input and a paired sampling strategy in multi-modal baselines. Dependent on the time-alignment, the RGB set $\mathcal{X}_{rgb}$ and the IR set $\mathcal{X}_{ir}$ are pair-sampled to form an image pair set, i.e., $\mathcal{X}_{pair} = \{(x_r^1, x_i^1), (x_r^2, x_i^2), \ldots, (x_r^n, x_i^n)\}$. To guarantee training robustness and accelerate parameter convergence, random shuffling is adopted to disrupt the order of image pairs. The shuffled image pairs are fed into the multi-modal model as dual-stream input.

Evaluation Metrics. The evaluation of detectors and localizers on the non-registered multi-modal track is still based on the metrics described in Section 4.1. Unlike the mono-modal track, the model in this track takes multi-modal data as dual-stream input and outputs paired results. Therefore, three models are designed for evaluation: `RGB-test`, `IR-test`, and `multi-test`. `RGB-test` and `IR-test` represent the evaluation of the RGB and IR test sets, respectively. `Multi-test` is simply combining the results of the RGB and IR test sets and evaluating them together without using additional data. We report the `multi-test` in order to consider both modalities' performance simultaneously and to facilitate comparisons between models.

4.3. Advantages and Challenges of Non-Registration

A common assumption of conventional multi-modal tasks is that the image pairs are well registered. However, performing the registration is a cumbersome operation that requires sophisticated sensors and heavy human efforts. Specifically, multi-modal image registration suffers from the following major problems:

(1) Multi-modal images often have different resolutions and fields of view due to the different sensors in the cameras. Therefore, it is almost impossible to achieve image registration simply by calibrating the inner parameters. It is also difficult to ensure that the image acquisition process is free from external disturbances even after the parameter calibration.

(2) (Automatic feature matching methods may not work well on data with complicated scenes and across modalities. A typical example is shown in Figure 4, where we use SURF [51] to achieve feature point extraction and matching. Our dataset cannot rely on such unsatisfactory matching results for automatic registration. Further, for tiny object analysis, matching key points for registration will be more difficult due to the low signal-to-noise ratio.

(3) The above two problems illustrate the difficulty of achieving the automatic alignment of multi-modal images. Therefore, the existing multi-modal dataset [3,21,46] almost relies on manual registration that requires processing each image pair and relies on a large amount of manpower. In addition, research shows that in practice, even well-processed data still suffer from weak alignment problems [36] (i.e., position shifts), which degrades model performance.

(4) Our dataset focuses on tiny objects at a distance. In this case, registration becomes more complicated since long-distance shots amplify small deviations. In scenes with dense and minute-scale objects, slight positional offsets may confuse the correspondence between objects in different modes.

In summary, multi-modal image registration is complex and not guaranteed to be completely accurate, and non-registration is an unavoidable problem.

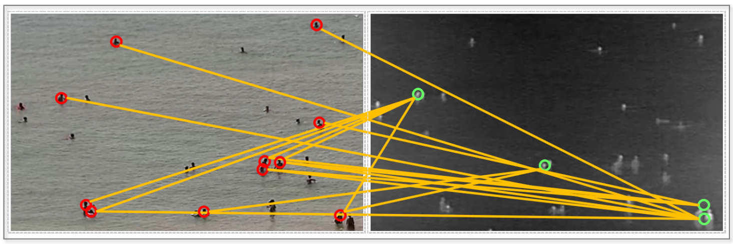

Figure 4. The poor feature-matching results suggest that, using our NRPerson dataset, it is difficult to achieve automatic registration.The red color represents the target in the visible image, the green color represents the target in the infrared image, and the yellow line indicates the matching between the two.

The above problems undoubtedly limit the development and application of multi-modal technology. Meanwhile, they confirm the necessity of introducing non-registration issues. The non-registered multi-modal track attempts to address these shortcomings by omitting the registration operation and only requiring the multi-modal image pairs to be time-aligned and spatial-overlapped. This not only saves time and manpower for registration but also performs data post-processing and object detection/localization in an end-to-end manner, greatly facilitating the generalization of multi-modal tasks to practical applications. However, the omission of registration results in the lack of instance-level correspondence established between image pairs. Therefore, using the multi-modal information effectively is challenging to improve detection performance.

5. Experiments

5.1. Experimental Settings

Our codes are based on MMDetection [52], and we use 8 RTX 3090 GPUs for training and evaluation.

Image cutting and pairing. Due to the limited memory of GPUs, we crop the original images into sub-images with overlaps during training and inference. We resize images of different resolutions to the same size (i.e., 1920 × 1080 pixels) and cut out a series of 640 × 640 patches. For the non-registered multi-modal track, multi-modal images are input in pairs, and thus the cropped sub-images need to be paired one by one. Sub-images at the corresponding positions in the original multi-modal image pair are formed into a new image pair, and the new paired images can be fed into the multi-modal network. According to previous works [48,49,53,54], pure background images (i.e., images without objects) are not used for training. For an image pair, there may be a situation where one image has objects, and the other does not. In our experiment, we filter out the image pairs in this situation and keep the pairs that both contain objects.

5.1.1. Mono-Modal Tiny Person Detection

In this detection track, Faster R-CNN [48], (Adaptive) RetinaNet [55], (Adaptive) RepPoints [56], Sparse R-CNN [57], Deformable DETR [58] and Swin-T (Faster R-CNN) [59] are included as the detectors. Faster R-CNN is based on the feature pyramid network (FPN) [16] with a ResNet-50 [60] backbone network. The number of training epochs is set to 12, and the stochastic gradient descent (SGD) optimizer is used with a learning rate of 0.01, decayed by 0.1 at the 8th and 11th epochs, respectively. Anchor sizes are set to [8, 16, 32, 64, 128], and aspect ratios are set to [0.5, 1.0, 2.0]. Since some images have dense objects in NRPerson, the maximum number of detection results per image is set to 1000 for evaluation.

To increase the data diversity and model robustness, we adopt various data augmentation methods, including random flipping, random cropping, and photometric distortion. There are eight kinds of transformations for photometric distortion: random brightness, random contrast, convert color from BGR to HSV, random saturation, random hue, convert color from HSV to BGR, random contrast, and randomly swap channels. In this paper, this series of data augmentation methods are identified as *standard augmentation*.

RetinaNet and RepPoints use the same experimental settings as Faster R-CNN and also adopt the standard augmentation.

Adaptive RetinaNet and Adaptive RepPoints denote the models with FPN adjustments for RetinaNet and RepPoints, respectively, which use the $(P_2, P_3, P_4, P_5, P_6)$ of the FPN instead of the default $(P_3, P_4, P_5, P_6, P_7)$. The purpose of this adjustment is to use features at a lower layer, which is more friendly for detecting tiny objects. The rest of the experimental settings are kept the same as RetinaNet and RepPoints.

Sparse R-CNN sets the experimental super-parameters somewhat differently from the above models. The number of training epochs is set to 36; the Adam optimizer is used with a learning rate of 2.5×10^{-5}, decayed by 0.1 at the 27th and 33rd epochs, respectively; and the number of learnable proposal boxes is set to 1000. In addition, data augmentation includes random scale jittering and random flipping.

Deformable DETR sets the number of training epochs to 50, uses the Adam optimizer with a learning rate of 1×10^{-4}, decayed by 0.1 at the 40th epoch and sets the number of queries to 1000. According to the paper, random scale jittering, random flipping, and random cropping augmentation are used.

Swin-T is a strong backbone which performs well in a broad range of vision tasks. We conduct our experiments with it on Faster R-CNN and keep the same settings as Faster R-CNN above except for the Adam optimizer with a learning rate of 1×10^{-4} and the 50 training epochs.

5.1.2. Mono-Modal Tiny Person Localization

In this localization track, P2PNet [12] is included as the localizer, and CPR [10] acts as the refinement module to optimize coarse point annotations. In addition, we convert the point-to-point problem in tiny person localization into a box-to-box one as described in Section 4.1. In this way, the above-mentioned detectors can be benchmarked in the localization track to explore the performance on multiple tasks.

P2PNet is specifically designed for point-based object localization. We follow the improvements to P2PNet in [10]: (1) We adopt ResNet-50 as the backbone network with P_2 of FPN as the output features. (2) The focal loss [55] is used when optimizing classification to better deal with the problem of imbalance. (3) The smooth-$\mathcal{L}_1$ loss [61] is used for regression. (4) We assign top-k positive samples for each ground-truth and regard the remaining samples as background in label assignment, and then perform non-maximum suppression (NMS) [62] on the points to obtain the final results. The number of training epochs of P2PNet is set as 12, and the learning rate is set as 1×10^{-4}, decayed by 0.1 at the 8th and 11th epochs, respectively.

CPR is a refinement module used to reduce the semantic variance during annotation. Hence, it can be combined with any localizer to provide refined point annotations. For CPR, the number of training epochs is set as 12, and the learning rate is set as 1×10^{-3}, decayed by 0.1 at the 8th and 11th epochs, respectively. The sampling radius R is set as 5.

Other localizers are converted from detectors in the detection track. They are trained with pseudo boxes refined and generated by CPR and adopt the same experimental settings as in the detection track.

5.1.3. Result Analysis

The detection and localization results of the mono-modal track are summarized in Table 4.

Table 4. The detection and localization results of representative models on the mono-modal track. For localization, the models adopt the annotations refined by CPR except for the explicit description of P2PNet. The **first**, **second** and **third** best results are marked in red, blue and black bold respectively.

Detection	**RGB-Modal**				**IR-Modal**			
	$\mathbf{AP_{50}}$	$\mathbf{AP^r_{50}}$	$\mathbf{AP^s_{50}}$	$\mathbf{AP^t_{50}}$	$\mathbf{AP_{50}}$	$\mathbf{AP^r_{50}}$	$\mathbf{AP^s_{50}}$	$\mathbf{AP^t_{50}}$
Faster R-CNN [48]	64.7	75.8	74.3	57.1	51.1	66.7	60.6	**44.5**
RetinaNet [55]	64.4	74.9	74.7	57.5	49.3	66.0	62.3	39.3
Adaptive RetinaNet [55]	65.5	74.1	74.3	**61.2**	**51.4**	66.5	**62.5**	**44.6**
RepPoints [56]	**66.8**	74.5	**76.7**	**60.3**	**51.4**	**67.9**	**64.2**	40.6
Adaptive RepPoints [56]	**67.1**	**76.1**	76.3	**63.5**	**51.9**	**67.2**	**64.8**	42.9
Sparse R-CNN [57]	66.4	**77.3**	76.1	53.7	48.1	63.2	59.3	**43.9**
Deformable DETR [58]	65.9	75.1	**77.9**	55.2	48.5	63.6	58.8	42.8
Swin-T (Faster R-CNN) [59]	**67.2**	**75.8**	**77.1**	57.7	49.2	**67.9**	58.8	42.6

Localization	**RGB-Modal**				**IR-Modal**			
	$\mathbf{AP_{1.0}}$	$\mathbf{AP^r_{1.0}}$	$\mathbf{AP^s_{1.0}}$	$\mathbf{AP^t_{1.0}}$	$\mathbf{AP_{1.0}}$	$\mathbf{AP^r_{1.0}}$	$\mathbf{AP^s_{1.0}}$	$\mathbf{AP^t_{1.0}}$
Faster R-CNN [48]	71.5	21.1	56.2	77.5	74.8	33.8	66.9	72.3
RetinaNet [55]	76.0	27.1	64.4	80.3	75.5	37.9	69.8	71.8
Adaptive RetinaNet [55]	74.5	25.6	61.5	79.5	74.7	35.3	67.3	71.8
RepPoints [56]	76.2	26.1	64.4	**81.2**	**76.5**	35.1	69.8	**73.7**
Adaptive RepPoints [56]	75.1	25.9	63.2	80.4	74.5	33.0	66.8	71.8
Sparse R-CNN [57]	**76.4**	**27.8**	**65.3**	80.2	74.2	34.0	68.1	70.1
Deformable DETR [58]	**78.0**	**29.7**	**68.2**	**83.2**	**77.7**	**40.9**	**73.1**	**75.8**
Swin-T (Faster R-CNN) [59]	72.7	23.2	58.4	76.9	76.3	**41.6**	**71.1**	**72.6**
P2PNet [12] *w/o* CPR [10]	71.4	25.9	51.9	**82.1**	66.4	26.3	54.2	72.0
P2PNet [12] *w*/CPR [10]	**80.1**	**33.6**	**71.4**	77.9	**81.5**	**46.5**	**77.6**	72.6

Detection. According to Figure 2, there are a large number of small-scale objects in NRPerson, but the detection performance for these *tiny* objects (in terms of AP^t_{50}) is not satisfactory, indicating that our dataset is a challenging benchmark for the state-of-the-art person detection algorithms. For objects with small scales, the spatial information of the network may be more important than the deeper layers. Therefore, Adaptive RetinaNet and Adaptive RepPoints improve performance (+1.7 and +1.0 AP_{50}, respectively) by adopting a lower layer of FPN to obtain more spatial information about objects. In particular, the detection performance for objects whose scales are *tiny* is significantly improved.

Localization. It can be seen that after the refinement coarse annotations by CPR, the models achieve highly competitive localization performance on our dataset. Different from the detection results, better localization performance is achieved on smaller objects, which indicates that small objects are not sensitive to the semantic variance [10] caused by coarse point annotations. Moreover, a *reasonable* object may be predicted with multiple results due to its large area, whereas a tiny object is an opposite. Consequently, there are fewer false positives on the small objects. The impressive performance on small objects demonstrates the significance and the application value of the localization task, especially in scenarios in which only the location of the object rather than the scale is concerned.

5.2. Non-Registered Multi-Modal Track

Multi-Modal Baseline. Since the non-registered multi-modal track is new, we construct a diverse set of natural multi-modal baselines. The multi-modal baselines employ the same models and experimental settings as in the mono-modal track and are simultaneously trained with RGB-modal and IR-modal data to explore adaptation to multi-modal detection and localization. Moreover, a comparison of multi-modal baselines with the mono-modal track can be used to explore the effect of adding additional data from other modalities on performance.

Result Analysis. The detection and localization results of the non-registered multi-modal track are summarized in Table 5. Although the correlation information between different modalities is not intentionally used in the multi-modal baselines, adding data from other modalities for training leads to the improved performance of many models (compared to those in the mono-modal track), especially in the RGB modality. Such performance improvement is reasonable and shows that if the complementary information between the multi-modal data is fully exploited, it will have a more positive impact on detection performance. The results provide sound empirical evidence on the superiority of multi-modal baselines in terms of their accuracy and applicability. However, there are also some models that show a degradation in performance, suggesting that simply adding multi-modal data does not always have a positive effect. It indicates the need to explore more effective fusion methods to fully utilize multi-modal information in future studies.

Table 5. The detection and localization results of the representative models on the non-registered multi-modal track. For localization, the models adopt the annotations refined by CPR except for the explicit description of P2PNet. The **first**, **second** and **third** best results are marked in red, blue and black bold, respectively.

Detection	**Multi-Test**				**RGB-Test**				**IR-Test**			
	AP$_{50}$	**AP$^{r}_{50}$**	**AP$^{s}_{50}$**	**AP$^{t}_{50}$**	**AP$_{50}$**	**AP$^{r}_{50}$**	**AP$^{s}_{50}$**	**AP$^{t}_{50}$**	**AP$_{50}$**	**AP$^{r}_{50}$**	**AP$^{s}_{50}$**	**AP$^{t}_{50}$**
Faster R-CNN [48]	60.1	**71.8**	70.4	**52.0**	66.5	77.3	77.1	55.1	50.9	**66.2**	57.6	35.9
RetinaNet [55]	58.9	70.3	**72.0**	49.6	65.4	74.6	76.9	53.0	49.4	65.0	58.5	32.0
Adaptive RetinaNet [55]	**60.6**	70.6	71.5	**53.8**	67.0	74.8	76.6	**56.8**	**51.1**	64.8	58.1	36.7
RepPoints [56]	**61.6**	**73.0**	**73.3**	**52.7**	66.4	72.7	74.2	**60.1**	**52.4**	**65.1**	**60.2**	36.0
Adaptive RepPoints [56]	**62.6**	**72.3**	**72.0**	50.4	**69.6**	**79.4**	**79.4**	**58.8**	**52.7**	**67.4**	**61.1**	35.7
Sparse R-CNN [57]	59.8	70.5	71.0	51.4	67.1	**78.4**	77.4	53.8	48.8	61.3	57.0	**37.7**
Deformable DETR [58]	60.1	71.3	71.4	51.7	**67.9**	**78.3**	**78.5**	54.4	49.3	63.9	57.8	**36.9**
Swin-T (Faster R-CNN) [59]	59.1	69.6	70.1	49.7	**68.6**	77.2	**77.9**	56.0	47.1	64.2	**58.9**	**38.5**
Localization	**Multi-Test**				**RGB-Test**				**IR-Test**			
	AP$_{1.0}$	**AP$^{r}_{1.0}$**	**AP$^{s}_{1.0}$**	**AP$^{t}_{1.0}$**	**AP$_{1.0}$**	**AP$^{r}_{1.0}$**	**AP$^{s}_{1.0}$**	**AP$^{t}_{1.0}$**	**AP$_{1.0}$**	**AP$^{r}_{1.0}$**	**AP$^{s}_{1.0}$**	**AP$^{t}_{1.0}$**
Faster R-CNN [48]	73.9	30.6	63.3	76.0	73.5	26.3	60.8	79.2	74.4	33.7	66.5	71.8
RetinaNet [55]	76.3	**34.6**	66.9	**77.8**	77.0	**31.8**	66.1	**81.3**	75.5	37.5	69.4	72.2
Adaptive RetinaNet [55]	75.5	33.2	65.5	77.1	75.9	29.9	64.1	80.7	75.0	36.4	68.6	71.5
RepPoints [56]	75.6	29.8	65.1	77.0	76.2	28.5	64.8	80.8	75.9	33.6	67.5	**73.3**
Adaptive RepPoints [56]	74.5	30.3	63.8	77.0	74.7	27.0	62.2	80.5	74.5	33.5	67.0	71.5
Sparse R-CNN [57]	**78.4**	34.4	**70.1**	**78.8**	**80.1**	**32.9**	**70.9**	**83.2**	**76.6**	35.9	**70.1**	**72.7**
Deformable DETR [58]	**77.1**	**34.5**	**69.5**	**80.5**	**77.5**	30.0	**68.3**	**83.2**	**77.7**	**39.6**	**73.0**	**76.1**
Swin-T (Faster R-CNN) [59]	72.6	33.3	63.0	75.0	72.0	24.0	58.2	77.2	73.3	**42.4**	68.3	71.3
P2PNet [12] *w/o* CPR [10]	66.2	24.0	48.9	75.0	68.9	25.8	49.4	80.4	63.4	22.9	49.2	67.2
P2PNet [12] *w*/CPR [10]	**80.3**	**39.4**	**74.4**	76.0	**80.8**	**35.3**	**73.8**	79.6	**78.6**	**43.8**	**74.3**	69.8

Visualization. We select some qualitative results for the visualization of Adap-RepPoints (with the best detection performance) and P2PNet (with the best localization performance) as shown in Figure 5. Since the size of objects in NRPerson is relatively small, we partially zoom in on the original images for display. The visualization results reflect that the models generally exhibit robust detection and localization capabilities. However, the models still have missing or wrong results for some areas where the objects are dense or blurred. This shows that our dataset is exceptionally challenging and needs to be continuously improved by follow-up research.

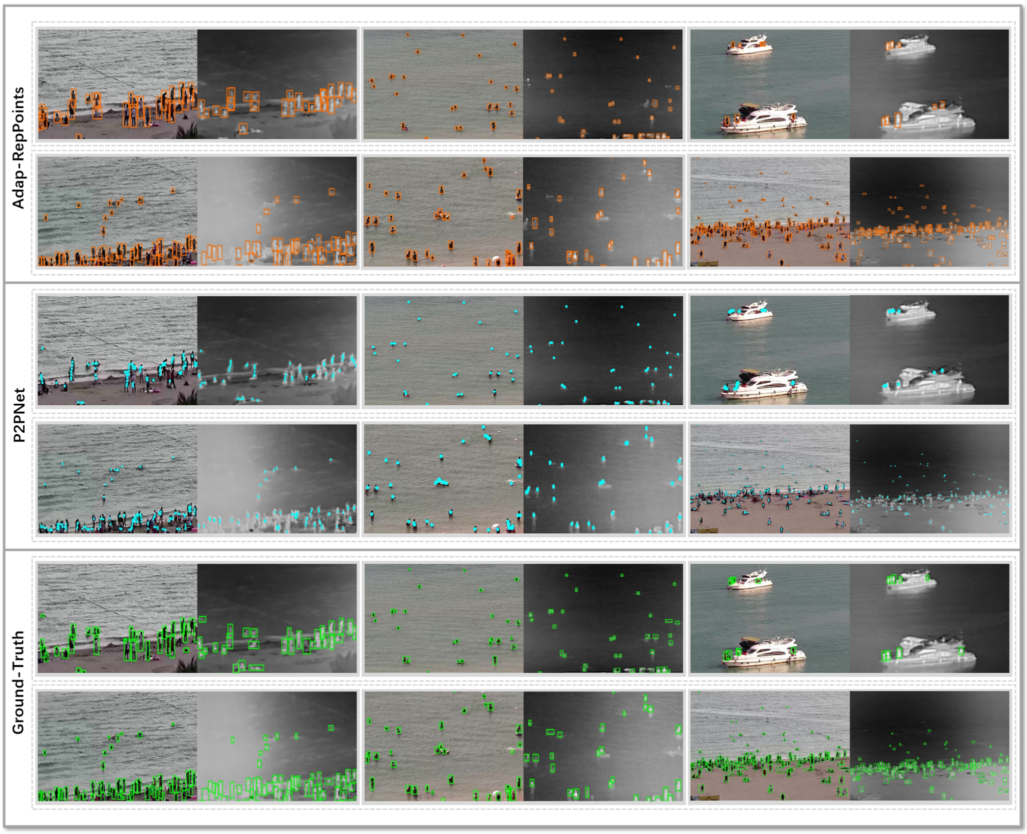

Figure 5. Some qualitative results of multi-modal baselines on the non-registered multi-modal track. The first two rows are the detection results of Adap-RepPoints (orange); the middle two rows are the localization results of P2PNet with CPR (blue); and the last two rows are the ground-truth (green). For better visualization results, we zoom in on the original images and capture the sub-images in which each pair has a roughly consistent field of view.

6. Conclusions

This paper contributes a non-registered multi-modal benchmark for tiny person detection and localization, called NRPerson. It introduces the concept of non-registration into the multi-modal task, eliminating the requirement of multi-modal image registration based on sophisticated sensors and heavy human efforts. We benchmark the representative models on the mono-modal track and the newly defined non-registered multi-modal track, benefiting from the empirical results to understand the detection/localization ability of different models and the impact of using multi-modal data. We construct a diverse set of natural multi-modal baselines, attempt to leverage non-registered multi-modal data for multi-task with a unified framework, and lay a robust foundation for facilitating future research.

We are committed to continuing our research using the NRPerson benchmark. Our future work will focus on exploring the detection of weak and small targets and addressing challenges of non-alignment in pretraining scenarios. Additionally, we plan to tackle the registration issues in non-aligned multimodal data, aiming to enhance the applicability and accuracy of detection methods under these complex conditions.

While our dataset is challenging, we believe that the baselines provided in this paper are far from reaching the upper limit of performance. Therefore, how to better utilize non-registered multi-modal data to deal with tiny objects is worth exploring. In the future, we will continue investigating more effective and robust models for the two tracks. We hope NRPerson can pave the way for follow-up studies in this area.

Author Contributions: Conceptualization, Y.Y. and X.H.; methodology, Y.Y. and X.H.; software, Y.Y., X.H. and K.W.; validation, Y.Y., X.H. and K.W.; formal analysis, Y.Y., X.H. and X.Y.; investigation, Y.Y. and X.Y.; resources, Y.Y., Z.H. and J.J.; data curation, Y.Y. and K.W.; writing—original draft

preparation, Y.Y. and X.H.; writing—review and editing, Y.Y., Z.W. and W.Y.; visualization, Y.Y., Z.W. and W.Y.; supervision, Z.H., G.L. and J.J.; project administration, Z.H., G.L. and J.J.; funding acquisition, Z.H. and J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data supporting the results reported in this study are openly available in a public repository. The datasets analyzed during the current study are available at the following GitHub repository: https://github.com/NRPerson/NRPerson. This repository includes all the datasets that were analyzed or generated during the course of this study. Further details and access to the data can be found directly via the provided URL. Researchers who wish to use these datasets are encouraged to cite this article and reference the data source appropriately in their work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: A benchmark. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
2. Zhang, S.; Benenson, R.; Schiele, B. Citypersons: A diverse dataset for pedestrian detection. In Proceedings of the In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
3. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A Visible-infrared Paired Dataset for Low-light Vision. *arXiv* **2021**, arXiv:2108.10831.
4. Zhang, Y.; Bai, Y.; Ding, M.; Xu, S.; Ghanem, B. KGSNet: Key-Point-Guided Super-Resolution Network for Pedestrian Detection in the Wild. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2251–2265. [CrossRef] [PubMed]
5. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
6. Torabi, A.; Massé, G.; Bilodeau, G.A. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. *Comput. Vis. Image Underst.* **2012**, *116*, 210–221. [CrossRef]
7. Wu, Z.; Fuller, N.; Theriault, D.; Betke, M. A thermal infrared video benchmark for visual analysis. In Proceedings of the CVPRW, Columbus, OH, USA, 23–28 June 2014.
8. FLIR, T. Free Flir Thermal Dataset for Algorithm Training. 2018. Available online: https://www.flir.com/oem/adas/adas-dataset-form/ (accessed on 29 March 2024).
9. Yu, X.; Gong, Y.; Jiang, N.; Ye, Q.; Han, Z. Scale match for tiny person detection. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020.
10. Yu, X.; Chen, P.; Wu, D.; Hassan, N.; Li, G.; Yan, J.; Shi, H.; Ye, Q.; Han, Z. Object Localization under Single Coarse Point Supervision. In Proceedings of the CVPR, New Orleans, LA, USA, 18–24 June 2022.
11. Yang, S.; Luo, P.; Loy, C.C.; Tang, X. Wider face: A face detection benchmark. In Proceedings of the CVPR, Las Vegas, NV, USA, 17–30 June 2016.
12. Song, Q.; Wang, C.; Jiang, Z.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Wu, Y. Rethinking counting and localization in crowds: A purely point-based framework. In Proceedings of the ICCV, Montreal, QC, Canada, 10–17 October 2021.
13. Gong, Y.; Yu, X.; Ding, Y.; Peng, X.; Zhao, J.; Han, Z. Effective fusion factor in FPN for tiny object detection. In Proceedings of the WACV, Waikoloa, HI, USA, 3–8 January 2021.
14. Jiang, N.; Yu, X.; Peng, X.; Gong, Y.; Han, Z. SM+: Refined Scale Match for Tiny Person Detection. In Proceedings of the ICASSP, Toronto, ON, Canada, 6–11 June 2021.
15. Yu, X.; Han, Z.; Gong, Y.; Jan, N.; Zhao, J.; Ye, Q.; Chen, J.; Feng, Y.; Zhang, B.; Wang, X.; et al. The 1st tiny object detection challenge: Methods and results. In Proceedings of the ECCV, Glasgow, UK, 23–28 August 2020.
16. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017.
17. Singh, B.; Davis, L.S. An analysis of scale invariance in object detection snip. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–23 June 2018.
18. Cao, G.; Xie, X.; Yang, W.; Liao, Q.; Shi, G.; Wu, J. Feature-fused SSD: Fast detection for small objects. In Proceedings of the ICGIP, Chengdu, China, 12–24 December 2018.
19. Duan, K.; Du, D.; Qi, H.; Huang, Q. Detecting small objects using a channel-aware deconvolutional network. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 1639–1652. [CrossRef]
20. Hu, X.; Xu, X.; Xiao, Y.; Chen, H.; He, S.; Qin, J.; Heng, P.A. SINet: A scale-insensitive convolutional neural network for fast vehicle detection. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 1010–1019. [CrossRef]
21. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So Kweon, I. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
22. Li, C.; Zhao, N.; Lu, Y.; Zhu, C.; Tang, J. Weighted sparse representation regularized graph learning for RGB-T object tracking. In Proceedings of the ACM MM, Mountain View, CA, USA, 23–27 October 2017.

23. Li, C.; Liang, X.; Lu, Y.; Zhao, N.; Tang, J. RGB-T object tracking: Benchmark and baseline. *Pattern Recognit.* **2019**, *96*, 106977. [CrossRef]
24. Nguyen, D.T.; Hong, H.G.; Kim, K.W.; Park, K.R. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* **2017**, *17*, 605. [CrossRef] [PubMed]
25. Wu, A.; Zheng, W.S.; Gong, S.; Lai, J. RGB-IR person re-identification by cross-modality similarity preservation. *Int. J. Comput. Vis.* **2020**, *128*, 1765–1785. [CrossRef]
26. Jiang, N.; Wang, K.; Peng, X.; Yu, X.; Wang, Q.; Xing, J.; Li, G.; Zhao, J.; Guo, G.; Han, Z. Anti-UAV: A large multi-modal benchmark for UAV tracking. *arXiv* **2021**, arXiv:2101.08466.
27. Sun, Z.; Zhao, F. Counterfactual attention alignment for visible-infrared cross-modality person re-identification. *Pattern Recognit. Lett.* **2023**, *168*, 79–85. [CrossRef]
28. Luo, X.; Jiang, Y.; Wang, A.; Wang, J.; Zhang, Z.; Wu, X. Infrared and visible image fusion based on Multi-State contextual hidden Markov Model. *Pattern Recognit.* **2023**, *138*, 109431. [CrossRef]
29. Xu, M.; Tang, L.; Zhang, H.; Ma, J. Infrared and visible image fusion via parallel scene and texture learning. *Pattern Recognit.* **2022**, *132*, 108929. [CrossRef]
30. Dollár, P.; Appel, R.; Belongie, S.J.; Perona, P. Fast Feature Pyramids for Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1532–1545. [CrossRef] [PubMed]
31. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognit.* **2019**, *85*, 161–171. [CrossRef]
32. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. In Proceedings of the BMVC, York, UK, 19–22 September 2016.
33. Konig, D.; Adam, M.; Jarvers, C.; Layher, G.; Neumann, H.; Teutsch, M. Fully convolutional region proposal networks for multispectral person detection. In Proceedings of the CVPRW, Honolulu, HI, USA, 21–26 July 2017.
34. Xu, D.; Ouyang, W.; Ricci, E.; Wang, X.; Sebe, N. Learning cross-modal deep representations for robust pedestrian detection. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017.
35. Zhang, L.; Liu, Z.; Zhang, S.; Yang, X.; Qiao, H.; Huang, K.; Hussain, A. Cross-modality interactive attention network for multispectral pedestrian detection. *Inf. Fusion* **2019**, *50*, 20–29. [CrossRef]
36. Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z.; Liu, Z. Weakly aligned cross-modal learning for multispectral pedestrian detection. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019.
37. Chen, Y.T.; Shi, J.; Ye, Z.; Mertz, C.; Ramanan, D.; Kong, S. Multimodal object detection via probabilistic ensembling. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 139–158.
38. Xiao, J.; Guo, H.; Zhou, J.; Zhao, T.; Yu, Q.; Chen, Y.; Wang, Z. Tiny object detection with context enhancement and feature purification. *Expert Syst. Appl.* **2023**, *211*, 118665. [CrossRef]
39. Ribera, J.; Guera, D.; Chen, Y.; Delp, E.J. Locating objects without bounding boxes. In Proceedings of the CVPR, Long Beach, CA, USA, 15–20 June 2019.
40. Yu, X.; Chen, P.; Wang, K.; Han, X.; Li, G.; Han, Z.; Ye, Q.; Jiao, J. CPR++: Object Localization via Single Coarse Point Supervision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024** . [CrossRef] [PubMed]
41. Leykin, A.; Ran, Y.; Hammoud, R. Thermal-visible video fusion for moving target tracking and pedestrian classification. In Proceedings of the CVPR, Virtual, 19–25 June 2007.
42. Toet, A. TNO Image Fusion Dataset. 2014. Available online: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029/1 (accessed on 29 March 2024).
43. Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; Sun, J. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv* **2018**, arXiv:1805.00123.
44. Xu, Z.; Zhuang, J.; Liu, Q.; Zhou, J.; Peng, S. Benchmarking a large-scale FIR dataset for on-road pedestrian detection. *Infrared Phys. Technol.* **2019**, *96*, 199–208. [CrossRef]
45. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the ECCV, Zurich, Switzerland, 6–12 September 2014; Volume 8693, pp. 740–755.
46. González, A.; Fang, Z.; Socarras, Y.; Serrat, J.; Vázquez, D.; Xu, J.; López, A.M. Pedestrian detection at day/night time with visible and FIR cameras: A comparison. *Sensors* **2016**, *16*, 820. [CrossRef]
47. Yang, K.; Yau, J.; Fei-Fei, L.; Deng, J.; Russakovsky, O. A study of face obfuscation in imagenet. *arXiv* **2021**, arXiv:2103.06191.
48. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef] [PubMed]
49. Zhao, Z.; Zheng, P.; Xu, S.; Wu, X. Object Detection With Deep Learning: A Review. *IEEE Trans. Neural Networks Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef]
50. Chen, P.; Yu, X.; Han, X.; Hassan, N.; Wang, K.; Li, J.; Zhao, J.; Shi, H.; Han, Z.; Ye, Q. Point-to-Box Network for Accurate Object Detection via Single Point Supervision. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022.
51. Bay, H.; Tuytelaars, T.; Gool, L.V. Surf: Speeded up robust features. In Proceedings of the ECCV, Crete, Greece, 5–11 September 2006.

52. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
53. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016.
54. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019.
55. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the ICCV, Venice, Italy, 22–29 October 2017.
56. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. Reppoints: Point set representation for object detection. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019.
57. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In Proceedings of the CVPR, Nashville, TN, USA, 20–25 June 2021.
58. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the ICLR, Virtual Event, 3–7 May 2021.
59. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the ICCV, Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
60. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016.
61. Girshick, R. Fast r-cnn. In Proceedings of the 2015 ICCV, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
62. Neubeck, A.; Gool, L.V. Efficient Non-Maximum Suppression. In Proceedings of the ICPR, Hong Kong, China, 20–24 August 2006.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.