Camo-M3FD: A New Benchmark Dataset for Cross-Spectral Camouflaged Pedestrian Detection

Henry O. Velesaca[1,3], Andrea Mero[1,4], Guillermo A. Castillo[1], Angel D. Sappa[1,2]

[1]ESPOL Polytechnic University, Campus Gustavo Galindo, 090902, Guayaquil, Ecuador

[2]Computer Vision Center, Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona, Spain

[3]Software Engineering Department, Research Center for Information and Communication Technologies (CITIC-UGR), University of Granada, 18071, Granada, Spain

[4]Università della Svizzera Italiana, Via Giuseppe Buffi 13, 6900, Lugano, Switzerland

{hvelesac,anmero,guancast}@espol.edu.ec, sappa@ieee.org

Abstract

Pedestrian detection is fundamental to autonomous driving, robotics, and surveillance. Despite progress in deep learning, reliable identification remains challenging due to occlusions, cluttered backgrounds, and degraded visibility. While multispectral detection—combining visible and thermal sensors—mitigates poor visibility, the challenge of camouflaged pedestrians remains largely unexplored. Existing Camouflaged Object Detection (COD) benchmarks focus on biological species, leaving a gap in safety-critical human detection where targets blend into their surroundings. To address this, we introduce Camo-M3FD (derived from the M3FD dataset), a novel benchmark for cross-spectral camouflaged pedestrian detection, consisting of registered visible-thermal image pairs. The dataset is curated using quantitative metrics to ensure high foreground-background similarity. We provide high-quality pixel-level masks and establish a standardized evaluation framework using state-of-the-art COD models. Our results demonstrate that while thermal signals provide indispensable localization cues, multispectral fusion is essential for refining structural details. Camo-M3FD serves as a foundational resource for developing robust and safety-critical detection systems. The dataset is available on GitHub: https://cod-espol.github.io/Camo-M3FD/.

1. Introduction

Pedestrian detection has been fundamental to the development of autonomous driving, robotics, and modern surveillance. However, despite the remarkable progress of deep learning–based detectors, reliable pedestrian detection in real-world scenarios remains challenging due to occlusions, cluttered backgrounds, diverse human poses, scale variation, and—critically—visibility degradation caused by nighttime conditions and adverse weather.[5, 17, 35]

To mitigate poor visibility, multispectral sensing has become a common strategy, combining visible-spectrum cameras with thermal sensors to acquire aligned image pairs [34]. Works such as [20] have shown that thermal information provides complementary cues for pedestrians when visual appearance is weak (e.g., at night, at far range, under partial occlusion, or across challenging weather conditions).

Beyond "difficult visibility" in general, camouflage introduces an even harder challenge, since targets are visually designed (or naturally appear) to blend into their surroundings. COD explicitly addresses this regime by focusing on objects that merge with background appearance and therefore require more sophisticated strategies than conventional detection paradigms. Surveys such as [41, 45] highlight that widely used COD datasets predominantly feature animals, leaving only a small fraction of samples containing people. Moreover, part of those samples involve military-grade camouflage, and in [26] it can be observed that the ground-truth annotations in these military-camouflage datasets often cover not only the person but also weapons. This further reduces the fraction of images that are truly relevant for human-centered operational settings (e.g., surveillance, security, or search and rescue). As a result, camouflaged pedestrian detection remains under-explored—a critical gap given that pedestrians are safety-critical targets and failures can have severe real-world consequences.

This work introduces Camo-M3FD, a new benchmark

dataset for cross-spectral camouflaged pedestrian detection developed from registered visible–thermal image pairs. The dataset is curated using quantitative camouflage metrics that capture foreground–background similarity in color, appearance, and boundary consistency to specifically retain pedestrian instances with strong environmental blending. In addition, high-quality pixel-level semantic ground-truth masks are provided. To enable standardized and comprehensive evaluation, a suite of widely recognized COD metrics is adopted, encompassing structural similarity, boundary accuracy, and global alignment. Furthermore, performance is reported using various adaptive and threshold-agnostic measures derived from precision–recall curves to ensure a representative and robust assessment of model capabilities.

The manuscript is organized as follows. Section 2 introduces related work, recent SOTA COD techniques, and methods that address the problem of the COD approach. Section 3 presents the methodology to construct the dataset. Then, Section 4 shows the experimental results using different COD techniques. Finally, discussion and conclusions are given in Section 5 and Section 6 respectively.

2. Related Works

This section situates the Camo-M3FD benchmark within the broader evolution of pedestrian detection and camouflaged object analysis. It reviews classical human detection methodologies, from handcrafted descriptors to early multispectral integrations, alongside contemporary state-of-the-art COD techniques. By analyzing how edge modeling and cross-spectral fusion address environmental blending, the theoretical foundation for high-precision, cross-spectral camouflaged pedestrian detection is established.

2.1. Classical Pedestrian Detection Techniques

Early pedestrian detection approaches were largely organized around a multi-scale sliding-window paradigm, where window-wise descriptors are extracted, a binary classifier is applied, a dense search over position/scale is performed, and results are finalized with non-maximum suppression (NMS) (e.g., [15], [16]); moreover, [8] argues for its suitability in low- to medium-resolution scenarios where segmentation- or keypoint-based methods tend to fail.

A major shift occurred with gradient-based descriptors, where the Histogram of Oriented Gradients (HOG) combined with linear SVMs became a strong baseline for pedestrian detection due to its sensitivity to the structure of the human silhouette; additionally, texture cues such as Local Binary Patterns (LBP) were fused with HOG to improve robustness to illumination and pose changes [2, 6].

To better handle articulation and occlusion, part-based formulations emerged, ranging from supervised part detectors to Deformable Part Models (DPM) that rely on a multi-scale HOG pyramid; learning is formulated via latent SVM supported by hard-negative mining; moreover, variants incorporating dimensionality reduction through PCA and multi-scale refinements report performance gains on benchmarks such as Inria [6] and Caltech [7, 8]. However, these approaches typically incur higher computational cost and are less suitable for real-time deployment without significant approximations [2, 13].

The work in [20] is relevant to our setting because it couples a multispectral benchmark with a strong classical baseline. The authors introduced a color–thermal pedestrian dataset captured with beam splitter-based hardware to physically align the two image domains, and proposed multispectral extensions of ACF by incorporating thermal-derived channels alongside conventional color/gradient channels. Their analysis shows that thermal cues can be particularly useful in challenging cases, such as long-range pedestrians and partial occlusions, establishing a practical reference point for cross-spectral pedestrian detection.

2.2. Camouflaged Object Detection Techniques

This section reviews contemporary state-of-the-art (SoTA) techniques in COD, focusing on diverse strategies designed to overcome the challenges of low-contrast boundaries and environmental blending.

The integration of explicit edge modeling and sophisticated feature fusion mechanisms characterizes recent advancements. BASNet [30] utilizes a predict-and-refine framework with a hybrid loss; however, it relies on structural similarity (SSIM) for implicit edge enhancement rather than explicit supervision. Conversely, SINet [12] adopts a biologically inspired search-and-identification pipeline that leverages edge cues to localize targets. Similarly, EAMNet [32] and BGNet [4] incorporate dedicated parallel branches or guidance modules to model object contours directly, while DGNet [21] captures edges via gradient flow to detect subtle contrast variations in low-texture regions. For global-local feature integration, CTF-Net [44] combines CNN-based local features with Transformer-based global context to improve boundary precision. In contrast, C^2F-Net [3] employs cross-level context fusion to reinforce structural coherence.

The challenge of cross-spectral and domain-specific detection has led to more specialized architectures. AVNet [36] introduces a cross-spectral attention-vision model specifically designed for ecological conservation, effectively fusing visible and infrared information to identify camouflaged targets in complex natural habitats. In parallel, PCNet [43] targets plant-specific camouflage through multi-scale refinement, addressing irregular edges characteristic of vegetation. Finally, iterative approaches such as HitNet [19] and uncertainty-aware models such as OCENet [25] provide dynamic supervision of high-uncertainty re-

gions, thereby indirectly improving boundary clarity. Collectively, these methods represent a paradigm shift toward balancing explicit geometric guidance with implicit contextual refinement across diverse modality scenarios.

3. Materials & Methods

This section delineates the systematic methodology employed in the curation and refinement of the Camo-M3FD dataset, a specialized benchmark derived from the M3FD dataset [24] for camouflaged pedestrian detection. Figure 1 shows example images of the final dataset obtained.

3.1. Data Collection

The foundation of this work is the M3FD dataset [24], a state-of-the-art multispectral benchmark widely utilized for object detection tasks involving registered thermal and visible-spectrum imagery. While M3FD covers a broad range of urban scenarios (i.e., road, campus, street, harsh weather, disguise, haze, forest, and others) and different types of objects, the primary objective of Camo-M3FD is to address the specific challenge of camouflaged pedestrian detection, a critical yet underserved task in autonomous surveillance and security applications. To construct this subset, an exhaustive manual audit of the M3FD repository[1] is performed, isolating only those frames containing pedestrian instances. This focused selection ensures that the resulting dataset provides a rigorous testbed for detecting human targets that exhibit high degrees of visual blending with their environmental surroundings.

3.2. Camouflage Quantification

Following the selection of pedestrian-centric imagery and because M3FD only contains bounding boxes, high-precision mask annotations are conducted using the CVAT (Computer Vision Annotation Tool)[2]. After making annotations on all the images, a rigorous filtering step is implemented to quantitatively define the "camouflage level" of each instance, ensuring the dataset's integrity as a specialized benchmark.

The quantification of camouflage levels is grounded in the tripartite metric framework proposed by Lamdouar et al. [22], which evaluates the visual relationship between a target and its immediate surroundings through color, texture, and structural coherence. The first component, Color Receptive Similarity (S_{rf}^{Q}), assesses the spectral alignment by comparing color distributions and intensities between the foreground and background. The second, Texture/Boundary Similarity (S_{b}^{Q}), measures the continuity of spatial patterns and the absence of disruptive edge gradients, determining how well the object's surface patterns blend into the environmental context. Finally, the Combined Camouflage Score (S_{α}^{Q}) acts as a holistic descriptor by performing a weighted fusion of S_{rf} and S_b, providing a single robust value that indicates total cryptic efficacy.

The S_α score is calculated for every annotated instance in the initial pool. To establish a robust threshold that accounts for the inherent distribution of the data, the median and standard deviation of the population are utilized. An image is classified as "valid camouflaged data" only if the target meets the following statistical criterion:

$$S_{\alpha}^{Q} \geq \text{Median}(S_{\alpha}^{Q}) - \sigma(S_{\alpha}^{Q}), \tag{1}$$

where σ denotes the standard deviation. This approach allows for the filtration of outliers—specifically, highly conspicuous targets—while maintaining a diverse range of challenging, naturally camouflaged scenarios.

3.3. Data Statistics and Selection

To provide a comprehensive overview of the dataset's characteristics, the spatial and geometric properties of the annotations are analyzed. Figure 2 illustrates the spatial distribution of the centroids of the annotated GT masks, demonstrating a varied coverage across the image plane. Figure 3 presents the aspect ratio distribution of the GT masks, highlighting the diversity in pedestrian poses and scales captured. Finally, Figure 4 provides a qualitative comparison of the selection process, showcasing examples of accepted and rejected images alongside their RGB-Sobel edges, the GT mask edges, and camouflage scores to validate the efficacy of the filtering threshold.

The statistical analysis of the initial dataset yielded a median S_{rf}^{Q} of 0.4636 with a standard deviation (σ) of 0.2881, while the S_b exhibited a median of 0.6130 and a lower dispersion of $\sigma = 0.0818$. The primary filtering metric, the Combined Camouflage Score (S_α), yielded a median of 0.5292 and a standard deviation of 0.1669. Based on these population statistics, the acceptance threshold is established as $\text{Median}(S_{\alpha}^{Q}) - \sigma(S_{\alpha}^{Q})$, defining a rigorous accepted range of $[0.3623, 1.0000]$ for the final dataset inclusion.

The rigorous filtering process resulted in a final curated set of 614 valid RGB-T image pairs of camouflaged pedestrians. This refined dataset is partitioned into training, validation, and testing sets using an 80/10/10 ratio. Specifically, 492 image pairs are used for training, while 61 and 61 pairs are reserved for validation and testing, respectively, ensuring a balanced distribution for model development and unbiased performance evaluation.

3.4. Evaluated Architectures

To benchmark the Camo-M3FD dataset, a representative selection of state-of-the-art (SoTA) models was evaluated, encompassing diverse architectural strategies for Camouflaged Object Detection (COD). These models are catego-

[1] https://github.com/dlut-dimt/TarDAL
[2] https://www.cvat.ai/

Figure 1. Example images of the Camo-M3FD dataset. *(1st row)* Visible (RGB) images. *(2nd row)* Thermal images. *(3rd row)* Segmentation mask images of camouflaged objects.

Dataset	Source	Year	Scope	Type of images	# images
Chameleon [31]	-	2018	Animal	RGB	76
CAMO [23, 42]	CVIU	2019	Animal & others	RGB	1,250
COD10K [11, 12]	CVPR	2020	Animal & others	RGB	10,000
NC4K [27]	CVPR	2021	Animal & others	RGB	4,121
Camo-M3FD (Ours)	CVPR	2026	Pedestrian	RGB + Thermal	614

Table 1. COD datasets comparison.

Table 2. Distinctive characteristics of the evaluated SoTA COD techniques.

Technique	**Source**	**Source Type**	**Year**	**Image Size (px)**	**Backbone**	**#Param. (M)**
BASNet [30]	CVPR	Conference	2019	256×256	ResNet-34 [18]	87.06
SINet-v2 [12]	TPAMI	Journal	2021	352×352	Res2Net-50 [14]	24.93
BGNet [4]	IJCAI	Conference	2022	416×416	Res2Net-50 [14]	77.80
C^2F-Net [3]	TCSVT	Conference	2022	352×352	Res2Net-50 [14]	26.36
OCENet [25]	WACV	Conference	2022	352×352	ResNet-50 [18]	58.17
EAMNet [32]	ICME	Conference	2023	384×384	Res2Net-50 [14]	30.51
DGNet [21]	MIR	Journal	2023	352×352	EfficientNet [33]	8.30
HitNet [19]	AAAI	Conference	2023	352×352	PVTv2 [40]	25.73
PCNet [43]	arXiv	-	2024	352×352	PVTv2 [40]	27.66
CTF-Net [44]	CVIU	Journal	2025	384×384	PVTv2 [40]	64.48
AVNet [36]	VISAPP	Conference	2026	416×416	PVTv2 [40]	48.04

rized based on their approach to boundary preservation and feature integration. Explicit edge-modeling networks, such as SINet [12], EAMNet [32], BGNet [4], and DGNet [21], are utilized to assess the efficacy of dedicated modules in capturing subtle contrast variations and object contours.

In parallel, implicit refinement frameworks are evaluated, including those focusing on hybrid losses (BASNet [30]), iterative feedback (HitNet [19]), and uncertainty modeling (OCENet [25]). Global-local context integration is assessed through Transformer-based and fusion-centric models like CTF-Net [44], C^2F-Net [3], and CHNet [38]. Furthermore, specialized channel-interaction networks, notably ARNet [37] and ARNet-V2 [39], are included to measure the impact of sophisticated feature refinement. Finally, the benchmark incorporates domain-specific and multimodal architectures, such as the plant-targeted PCNet [43]

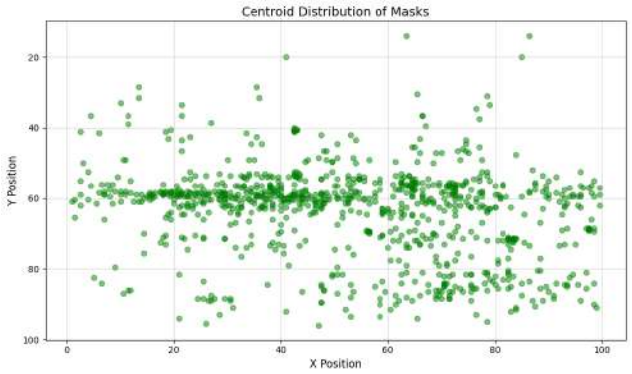

Figure 2. Spatial distribution of the centroids of the annotated GT masks.

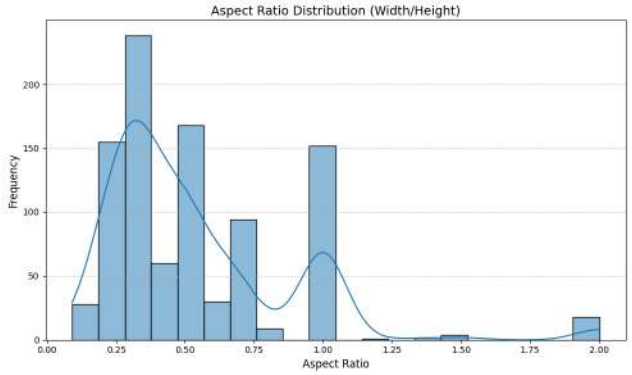

Figure 3. Aspect-ratio distribution of the GT masks.

and the cross-spectral AVNet [36], to evaluate performance across diverse ecological and multimodal scenarios. A comprehensive summary of these architectural characteristics is provided in Table 2.

3.5. Metrics

The performance of the SoTA COD approaches mentioned in the previous section is rigorously evaluated using five standard quantitative metrics, ensuring a holistic assessment of model accuracy and robustness. These metrics include: Structure-measure (S_α) [9], weighted F-measure (F_β^w) [28], Mean Absolute Error (M) [29], Enhanced-alignment measure (E_ϕ) [10], and the traditional F-measure (F_β) [1]. The S_α metric is utilized to quantify structural similarity by evaluating both region-aware and object-aware correlations between the prediction and ground truth, thereby measuring the preservation of global structural integrity. To address the limitations of pixel-wise comparisons, the F_β^w incorporates spatial weights to provide an improved assessment of segmentation quality, emphasizing boundary precision and the spatial distribution of errors. For pixel-level accuracy, the M metric calculates the average absolute difference between the normalized saliency maps and the binary ground truth. Furthermore, the E_ϕ metric leverages human visual perception mechanisms to simultaneously evaluate local pixel matching and global image statistics. Lastly, F_β offers a harmonic mean of precision and recall, serving as a fundamental measure of overall detection efficacy. To capture the performance across varying confidence levels, multiple variants of the F-measure and E-measure are computed. This includes the adaptive threshold version (F_β^{adp}, E_ϕ^{adp}), as well as the mean (F_β^{mean}, E_ϕ^{mean}) and maximum (F_β^{max}, E_ϕ^{max}) values derived from the precision-recall curves. This multi-faceted evaluation strategy ensures that the proposed framework is benchmarked against both deterministic and threshold-agnostic performance criteria.

4. Experimental Results

The experimental evaluation of the Camo-M3FD dataset utilizes a comprehensive suite of SoTA COD models to establish a robust performance baseline. This section analyzes the quantitative metrics across different modalities and provides a qualitative assessment of the segmentation challenges inherent in cross-spectral pedestrian camouflage.

4.1. Quantitative Evaluation

As summarized in Table 3, the quantitative results underscore the significant impact of spectral modality on detection accuracy. The quantitative evaluation across various state-of-the-art architectures reveals several critical trends regarding the performance of camouflaged object detection on the Camo-M3FD dataset. A consistent observation across all evaluated models is that the thermal modality significantly outperforms the visible baseline, suggesting that thermal signatures provide indispensable cues for localizing pedestrians when their visual appearance blends seamlessly into the background. Among the evaluated techniques, AVNet consistently achieves the highest performance, particularly when leveraging its native multispectral integration. It achieves the highest scores in nearly all key metrics, including a S_α of 0.7318, a weighted F-measure (F_β^w) of 0.5301, and an enhanced-alignment measure (E_ϕ^{mean}) of 0.8287. This underscores the effectiveness of cross-spectral fusion in resolving complex camouflage scenarios that a single modality cannot fully address.

Furthermore, models that incorporate explicit boundary modeling or uncertainty awareness, such as OCENet and BGNet, demonstrate high competitive resilience in the thermal domain, exhibiting strong structural alignment and precision. The error margins across the top-tier models remain notably low, indicating high spatial accuracy in the predicted masks even under challenging environmental blending. Overall, these results confirm that while single-modality models show a clear preference for thermal data, the most robust detection is achieved by effectively utilizing the complementary nature of both spectra.

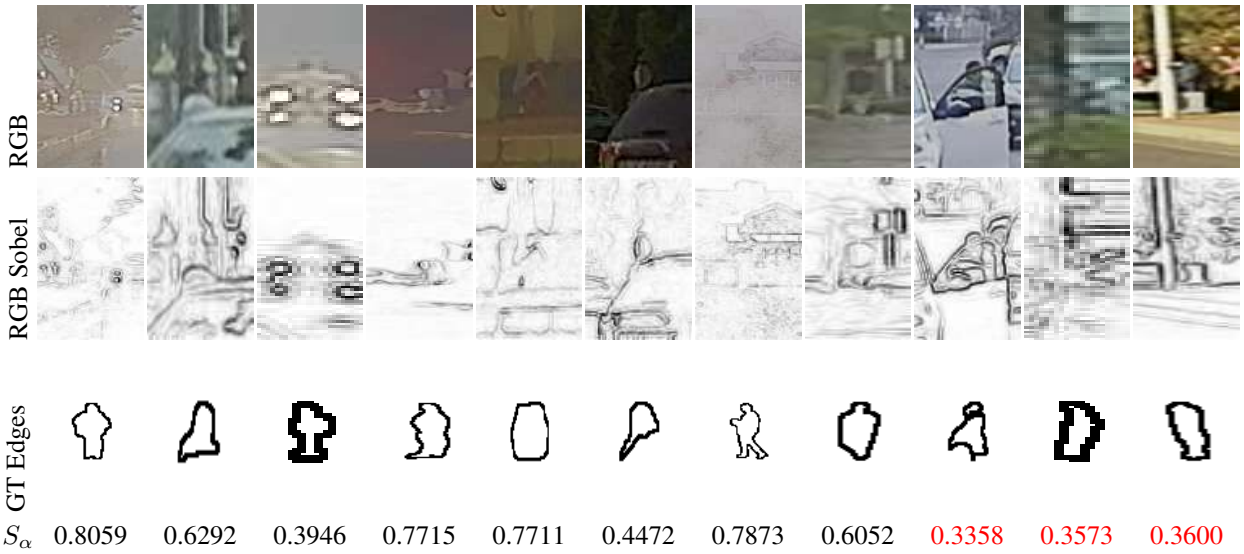

Figure 4. Examples of accepted and rejected (marked in red) images alongside their respective edges extracted by RGB using Sobel, edges of the GT mask, and camouflage scores (S_α).

Table 3. Metric evaluation results for each COD technique on the Camo-M3FD dataset, reported for the RGB and Thermal baseline. Results are presented using the metric notation defined in Sec. 3.5, "↑ / ↓" indicates that larger or smaller is better. The best three performing results are highlighted using color: First, Second, and Third respectively.

Technique	Input	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$	$E_\phi^{max} \uparrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$F_\beta^{max} \uparrow$
BASNet [30]	Vis	0.6239	0.2902	0.0032	0.6972	0.7183	0.8042	0.2879	0.3057	0.3137
	Th	0.7051	0.4161	0.0028	0.7293	0.7822	0.8078	0.3762	0.4358	0.4571
SINet-v2 [12]	Vis	0.6275	0.2693	0.0037	0.6039	0.7080	0.7227	0.2244	0.2872	0.3033
	Th	0.6927	0.4072	0.0034	0.6450	0.7593	0.7949	0.3428	0.4244	0.4424
BGNet [4]	Vis	0.6745	0.3922	0.0500	0.7594	0.7687	0.8142	0.3576	0.4124	0.4255
	Th	0.7196	0.4699	0.0106	0.7664	0.8306	0.8539	0.4315	0.4865	0.4963
C^2F-Net [3]	Vis	0.5137	0.0432	0.0811	0.4804	0.6079	0.7333	0.1433	0.2155	0.2554
	Th	0.5244	0.0522	0.0663	0.5122	0.6432	0.7656	0.2064	0.2882	0.3437
OCENet [25]	Vis	0.5994	0.2357	0.0037	0.6680	0.7975	0.8201	0.2240	0.2546	0.2632
	Th	0.7277	0.4884	0.0037	0.7122	0.8152	0.8666	0.4253	0.4998	0.5403
EAMNet [32]	Vis	0.5227	0.0494	0.0160	0.4048	0.6141	0.8109	0.0998	0.1752	0.2352
	Th	0.5047	0.0333	0.0506	0.4946	0.6458	0.8091	0.1836	0.2622	0.3799
DGNet [21]	Vis	0.6438	0.3109	0.0039	0.6720	0.7598	0.7739	0.2759	0.3235	0.3377
	Th	0.6898	0.4073	0.0052	0.6765	0.7928	0.8227	0.3586	0.4244	0.4403
HitNet [19]	Vis	0.5659	0.1593	0.0030	0.7333	0.5685	0.7353	0.1815	0.1721	0.1809
	Th	0.6682	0.3622	0.0029	0.7694	0.7466	0.7778	0.3910	0.3800	0.3919
PCNet [43]	Vis	0.6512	0.3227	0.0034	0.5048	0.7639	0.8069	0.1688	0.3464	0.3552
	Th	0.7034	0.4260	0.0030	0.6187	0.8280	0.8428	0.2674	0.4504	0.4572
CTF-Net [44]	Vis	0.5077	0.0525	0.0755	0.4201	0.5912	0.7296	0.1322	0.2449	0.3146
	Th	0.6532	0.2955	0.0116	0.4178	0.7515	0.8073	0.1409	0.4310	0.4794
AVNet [36]	Vis	0.6669	0.4035	0.0029	0.8294	0.8164	0.8287	0.3923	0.3985	0.4068
	Th	0.7289	0.5066	0.0026	0.7989	0.8075	0.8242	0.4831	0.4926	0.5113
	Vis+Th	0.7318	0.5301	0.0030	0.8167	0.8287	0.8617	0.5051	0.5139	0.5362

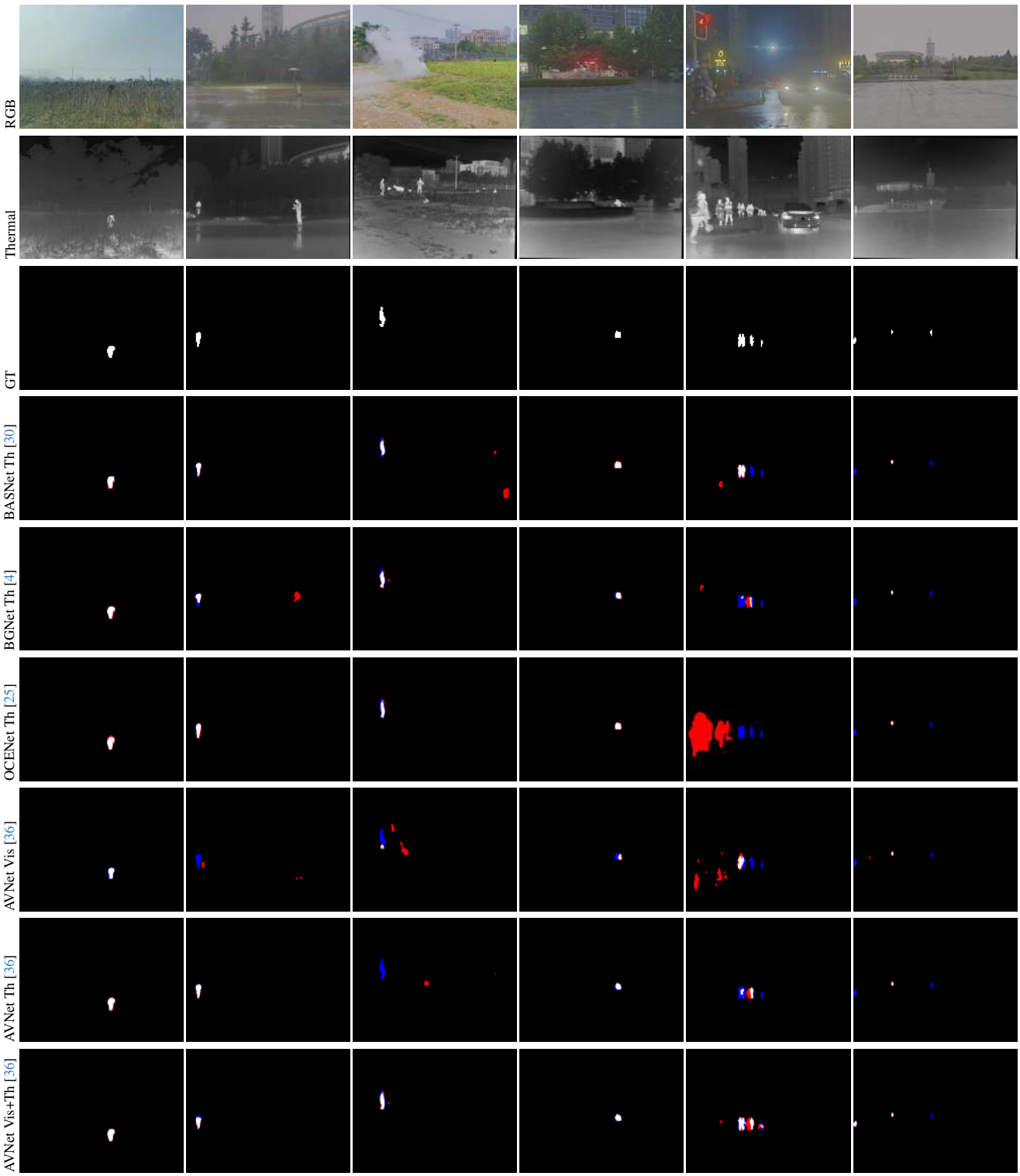

Figure 5. Results using SoTA COD techniques that have achieved first or second place in at least one of the metrics. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (blue areas, miss-segmentation).

4.2. Qualitative Evaluation

The qualitative analysis presented in Figure 5 reinforces the metric-based findings by visualizing the segmentation successes and failures of the leading models. A primary observation is the prevalence of "miss-segmentation" (blue areas) in visible-only or thermal-only predictions, where models fail to detect extremities or entire torso sections that blend seamlessly with the background texture. In contrast, predictions from thermal-based models and the multispectral AVNet exhibit much higher structural coherence, successfully capturing the human silhouette. However, the dataset presents a persistent challenge in the form of "over-segmentation" (red areas), where models erroneously include background elements with similar heat signatures or edge profiles as part of the pedestrian mask. The results from AVNet (Vis+Th) show the most accurate alignment with the ground truth (white areas), effectively using the visible channel to refine boundaries that might be blurred in the thermal domain. These results collectively validate Camo-M3FD as a challenging benchmark that requires sophisticated multimodal fusion to achieve high-precision pedestrian detection in complex environments.

5. Discussion

The results obtained from the proposed Camo-M3FD benchmark provide critical insights into the nature of pedestrian camouflage across different spectral domains. A primary finding is the inherent limitation of visible-spectrum (RGB) sensors when dealing with advanced camouflage, where the foreground-background similarity in texture and color often leads to significant miss-segmentation. Our experiments demonstrate that the integration of thermal imaging is not merely an enhancement but a necessity for reliable detection in these scenarios. The thermal modality effectively "breaks" the visual camouflage by highlighting heat signatures that are nearly impossible to mask with traditional visual techniques.

However, the "over-segmentation" observed in several high-performing models—where non-pedestrian heat sources are incorrectly classified—indicates that thermal data alone is not a definitive solution. This suggests that future research should focus on more sophisticated cross-modal interaction modules. Specifically, architectures that can dynamically weigh the importance of each spectrum based on environmental conditions (e.g., thermal noise in hot weather vs. visual clutter) are likely to define the next generation of COD models. Furthermore, the performance gap between existing SoTA models on our dataset compared to traditional COD benchmarks confirms that Camo-M3FD presents a more complex and realistic challenge, moving beyond static biological camouflage into dynamic, real-world pedestrian scenarios.

6. Conclusions

This paper introduces Camo-M3FD, a novel benchmark dataset specifically designed for cross-spectral camouflaged pedestrian detection. An extensive evaluation of state-of-the-art models establishes a clear baseline for future work, highlighting the superior performance of multispectral fusion approaches, such as AVNet, over single-modality methods. The findings emphasize that while thermal signals provide indispensable cues for localization, the visual domain remains crucial for refining structural details. It is expected that Camo-M3FD will serve as a foundational resource for the community, encouraging the development of more robust, safety-critical detection systems for autonomous driving and surveillance.

Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-24-1-0206; and partially supported by the Grant PID2021-128945NB-I00 funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF/EU and Grant PID2024-162815NB-I00 funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF/EU; and by the ESPOL project "Advancing Camouflaged Object Detection with a cost-effective Cross-Spectral vision system (ACODCS)" (CIDIS-003-2024). The authors acknowledge the support of the Generalitat de Catalunya CERCA Program to CVC's general activities.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009. 5

[2] Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta, and Vitoantonio Bevilacqua. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, 300:17–33, 2018. 2

[3] Geng Chen, Si-Jie Liu, Yu-Jia Sun, Ge-Peng Ji, Ya-Feng Wu, and Tao Zhou. Camouflaged object detection via context-aware cross-level fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6981–6993, 2022. 2, 4, 6

[4] Tianyou Chen, Jin Xiao, Xiaoguang Hu, Guofeng Zhang, and Shaojie Wang. Boundary-guided network for camouflaged object detection. *Knowledge-based systems*, 248: 108901, 2022. 2, 4, 6, 7

[5] Wei Chen, Yuxuan Zhu, Zijian Tian, Fan Zhang, and Minda Yao. Occlusion and multi-scale pedestrian detection a review. *Array*, 19:100318, 2023. 1

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer soci-*

ety conference on computer vision and pattern recognition (CVPR'05), pages 886–893. Ieee, 2005. 2

[7] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE conference on computer vision and pattern recognition*, pages 304–311. IEEE, 2009. 2

[8] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011. 2

[9] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017. 5

[10] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018. 5

[11] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, 2020. 4

[12] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6024–6042, 2021. 2, 4, 6

[13] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008. 2

[14] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019. 4

[15] David Gerónimo, Angel D Sappa, Antonio López, and Daniel Ponsa. Pedestrian detection using adaboost learning of features and vehicle pitch estimation. In *Proceedings of the International Conference on Visualization, Imaging, and Image Processing, Palma de Mallorca, Spain*, 2006. 2

[16] David Gerónimo, Antonio López, and Angel D Sappa. Computer vision approaches to pedestrian detection: visible spectrum survey. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 547–554. Springer, 2007. 2

[17] Bahareh Ghari, Ali Tourani, Asadollah Shahbahrami, and Georgi Gaydadjiev. Pedestrian detection in low-light conditions: A comprehensive survey. *Image and Vision Computing*, 148:105106, 2024. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conf. on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[19] Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Donghao Luo, Ying Tai, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 881–889, 2023. 2, 4, 6

[20] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 1, 2

[21] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 20(1):92–108, 2023. 2, 4, 6

[22] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. The making and breaking of camouflage. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 832–842, 2023. 3

[23] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Journal of Computer Vision and Image Understanding*, 184:45–56, 2019. 4

[24] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. 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*, pages 5802–5811, 2022. 3

[25] Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1445–1454, 2022. 2, 4, 6, 7

[26] Yang Liu, Cong-qing Wang, and Yong-jun Zhou. Camouflaged people detection based on a semi-supervised search identification network. *Defence Technology*, 21:176–183, 2023. 1

[27] Yunqiu Lyu, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Conf. on Computer Vision and Pattern Recognition*, 2021. 4

[28] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255, 2014. 5

[29] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE conference on computer vision and pattern recognition*, pages 733–740. IEEE, 2012. 5

[30] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4, 6, 7

[31] Przemysław Skurowski, Hassan Abdulameer, Jakub Błaszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018. 4

[32] Dongyue Sun, Shiyao Jiang, and Lin Qi. Edge-aware mirror network for camouflaged object detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2465–2470. IEEE, 2023. 2, 4, 6

[33] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International*

conference on machine learning, pages 6105–6114. PMLR, 2019. 4

[34] Michael Teutsch, Angel D Sappa, and Riad I Hammoud. Computer vision in the infrared spectrum: challenges and approaches. 2021. 1

[35] Paulius Tumas, Adam Nowosielski, and Arturas Serackis. Pedestrian detection in severe weather conditions. *Ieee Access*, 8:62775–62784, 2020. 1

[36] Henry Velesaca, Hector Villegas, and Angel Sappa. AVNet: Cross-Spectral Attention-Vision Model for Camouflaged Object Detection in Ecological Conservation. In *Proceedings of the 14th International Conference on Computer Vision Theory and Applications*, pages 1–10. INSTICC, SciTePress, 2026. 2, 4, 5, 6, 7

[37] Kuan Wang, Xiuhong Li, Yulong Bai, Songlin Li, Mengge Lu, and Zhenhong Jia. Assisted refinement network based on channel information interaction for camouflaged object detection. In *Int. Conf. on Multimedia Retrieval*, pages 2058–2062, 2025. 4

[38] Kuan Wang, Xiuhong Li, Songlin Li, Yulong Bai, Boyuan Li, Mengge Lu, and Zhenhong Jia. Efficient camouflaged object detection network based on channel reconstruction and hybrid attention. In *Int. Conf. on Multimedia Retrieval*, pages 2063–2067, 2025. 4

[39] Kuan Wang, Yanjun Qin, Mengge Lu, Liejun Wang, and Xiaoming Tao. Assisted refinement network based on channel information interaction for camouflaged and salient object detection. *arXiv preprint arXiv:2512.11369*, 2025. 4

[40] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 4

[41] Fengyang Xiao, Sujie Hu, Yuqi Shen, Chengyu Fang, Jinfa Huang, Chunming He, Longxiang Tang, Ziyun Yang, and Xiu Li. A survey of camouflaged object detection and beyond. *arXiv preprint arXiv:2408.14562*, 2024. 1

[42] Jinnan Yan, Trung-Nghia Le, Khanh-Duy Nguyen, Minh-Triet Tran, Thanh-Toan Do, and Tam V. Nguyen. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300, 2021. 4

[43] Jinyu Yang, Qingwei Wang, Feng Zheng, Peng Chen, Aleš Leonardis, and Deng-Ping Fan. Plantcamo: Plant camouflage detection. *arXiv preprint arXiv:2410.17598*, 2024. 2, 4, 6

[44] Dongdong Zhang, Chunping Wang, Huiying Wang, Qiang Fu, and Zhaorui Li. An effective cnn and transformer fusion network for camouflaged object detection. *Computer Vision and Image Understanding*, page 104431, 2025. 2, 4, 6

[45] Junmin Zhong, Anzhi Wang, Chunhong Ren, and Jintao Wu. A survey on deep learning-based camouflaged object detection. *Multimedia Systems*, 30(5):268, 2024. 1