Semantically Grounded Explainability for Pedestrian Detection Using Anatomical Proportions and Grid-Based Visualization

Chiara Galdi
dept. of Digital Security
EURECOM
Sophia Antipolis, France
chiara.galdi@eurecom.fr

Romain Giot
University of Bordeaux, CNRS,
Bordeaux INP, LaBRI, UMR 5800,
33400, Talence, France
romain.giot@u-bordeaux.fr

***Abstract*—Explainable computer vision aims to understand how and why complex vision systems make decisions. However, existing explainability methods often highlight relevant image regions without providing semantically meaningful explanations for human users. In this work, we analyze the behavior of a pedestrian detector under semantically meaningful image perturbations. To overcome the interpretability limitations of existing explainability approaches, pedestrian occlusions are defined according to anatomical proportions derived from Leonardo da Vinci's golden ratio, enabling a human-understandable decomposition of body regions. In addition, a grid-based visualization of the input and latent representations is employed to provide a structured global view of the detector behavior. Experimental results show that although the detector attention is strongly concentrated on the leg region, the largest performance degradation occurs when the upper body is occluded, with the Average Precision decreasing from 86.51% for full-body pedestrians to 35.25% under upper-body occlusion. These results suggest that pedestrian detectors rely on complementary semantic body cues beyond the most visually salient regions. The proposed framework provides complementary local and global explainability insights into pedestrian detection models. The code, visualizations, and experimental results are made publicly available at: https://github.com/eurecom-fscv/ExFMA_Semantically_Grounded_Explainability_for_Pedestrian_Detection**

I. Introduction

Trust is a key and central issue in computer vision (CV). The degree of trust on CV solutions is still an open challenge in the state-of-the-art and several examples have shown that CV systems exhibit various weaknesses and biases [15], [20]. This is in particular true for pedestrian detection (PD) based on machine learning [13], whose performance largely depends on training datasets and the exact approach used in training, to mention two among several problems that have been identified.

To address these limitations and improve the reliability of CV systems, it is essential to better understand how and why such models make their decisions. This challenge has led to the emergence of the field of Explainable Artificial Intelligence (XAI) [2], [11], which aims to provide interpretable representations of model behavior and to highlight the visual evidence used during inference.

This work was supported by the French Agence Nationale de la Recherche (ANR) via the GOOD-BIAS (ANR-25-CE39-6459) project and by the LaBRI.

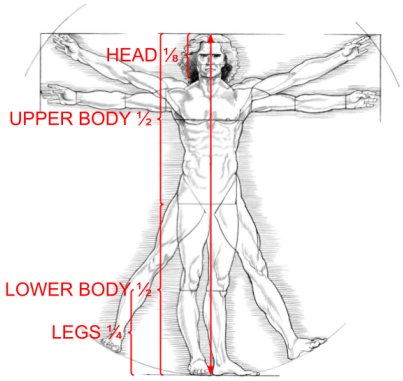


Fig. 1: Leonardo da Vinci's *Vitruvian Man*, illustrating classical human body proportions inspired by the golden ratio and used in this work to define semantically meaningful anatomical regions for explainability analysis.

XAI methods can generally be divided into two complementary categories: local XAI and global XAI. Local XAI aims to explain individual predictions by identifying the regions or features that contributed most to a specific decision, while global XAI seeks to provide a broader understanding of the overall behavior and structure learned by the model across a dataset. In CV, most existing methods focus primarily on local explanations [19], [22], [28]. Moreover, they often lack semantic interpretability, as they primarily indicate which parts of the image contribute to the decision without providing a clear, human-understandable explanation of how these regions relate to meaningful concepts or how they influence the final prediction [8], [27]. In this work, we propose a framework that combines both local and global XAI for PD. Local XAI is achieved through anatomically grounded occlusion analysis and Ablation-CAM visualizations, while global XAI is provided through structured grid-based visualization of the detector representations. By grounding the perturbations in

semantically meaningful human body regions derived from the golden ratio, the proposed approach provides explanations that are more interpretable from a human perspective.

The remainder of this paper is organized as follows. Section II reviews related work on XAI methods for CV and high-dimensional data visualization. Section III presents the proposed XAI framework, including the anatomically grounded body partitioning inspired by the golden ratio and the global visualization strategy. Section IV describes the experimental protocol, the pedestrian detection pipeline, and the XAI setup based on Ablation-CAM. Section V discusses the experimental results and analyzes the contribution of different body regions to pedestrian detection. Finally, Section VI concludes the paper and outlines future research directions.

II. Related Works

A. Explainability Methods for Computer Vision

Existing approaches can be broadly categorized into several families depending on how explanations are generated [17].

Perturbation-based methods, such as Local Interpretable Model-agnostic Explanations [22] (LIME) and Randomized Input Sampling for Explanation [19] (RISE), estimate the importance of image regions by observing the effect of systematically masking parts of the input. These approaches are model-agnostic and flexible. However, LIME's superpixel-based decomposition does not naturally align with the anatomically defined regions considered in this work. Similarly, the randomness of the masks in RISE makes it difficult to enforce semantically meaningful structures such as body-part regions.

Backpropagation-based methods explain model predictions by propagating information backward through the network from the output layer toward the input image in order to estimate the contribution of internal activations and input pixels to the final decision. Among them, Layer-wise Relevance Propagation [4] (LRP) redistributes the prediction score backward through the network to highlight relevant regions. These methods provide fine-grained explanations but can be sensitive to noise and architectural choices. The resulting explanations are often highly detailed but lack clear semantic aggregation, which limits their interpretability at the level of human body regions. Class Activation Maps [28], or CAMs, modify the network architecture by replacing fully-connected layers in the end by a Global Average Pooling layer and concatenates the averages of the activations of convolutional feature maps that are located before the final output layer and create a feature vector. The weighted sum of the vector is fed to the final softmax loss layer. Finally, the important image regions are identified by projecting back the weight of the output layer to the convolutional feature maps. The main drawback of CAM is that it requires neural networks to have a specific architecture in the final layer. This is not the case for the Grad-CAM [23] method, which is a generalization of CAM that can produce visual explanations for any CNN architecture. As a gradient-based method, Grad-CAM uses the class-specific gradient information flowing into the final convolutional layer of a CNN, producing a coarse localization map of the important regions in the image. Feature Map Explanation [3] (FEM), offer a simpler alternative by directly leveraging feature maps. FEM is computationally efficient and easy to implement, but may provide less precise localization compared to gradient-based approaches.

Finally, Ablation-CAM [9] provides an alternative that estimates feature importance through systematic perturbation of internal network representations. Instead of relying on gradients, Ablation-CAM measures the contribution of each feature map by observing the change in the model's output when that feature map is selectively ablated (i.e., set to zero). The importance weights are computed based on the drop in prediction confidence, and are then used to generate a class activation map in a manner similar to CAM. This approach combines the robustness of perturbation-based methods with the efficiency of activation-based techniques, while avoiding some of the noise and instability associated with gradient-based explanations.

In this work, we rely on Ablation-CAM to generate the XAI maps used in our analysis, as it provides stable and architecture-independent attributions that are well suited to object detection models.

B. Visualization of High Dimensional Data

It is common in the literature to visualize high dimensional data (and especially latent space of deep learning models) using dimensionality reduction techniques such as PCA [18], t-SNE [26], UMAP [14], etc. The dataset is projected from n to 2 dimensions, and then depicted using scatter plots. However, these techniques often fail to capture the complex relationships between data points and can lead to misleading visualizations: 2d neighborhoods are not systematically the same as high dimensional neighborhoods, the representation does not allow to get data density because of overlaps, screen space is not properly used with unused pixels, etc.

Grid-based visualization techniques have been proposed to address some of these issues. Although they do not always solve the neighborhood problem, they allow to get a better representation of the data density and to use the screen space more efficiently. Different families of grid arrangement methods exist, such as: sorting based methods (e.g. SSM [25]), geometric based methods (e.g. VRGRID [12]), manifold based methods (e.g. IsoMatch [10] or any method based on 2d dimension reduction followed by a grid assignment), filtering and swapping based methods (e.g. FLAS [7]), or gradient-descend based approaches (e.g. GradSort [6], or ShuffleSort [5]). The cost of these methods is often higher than dimensionality reduction techniques by orders of magnitude; however the visualization quality is often better, and this cost is not a problem for small datasets (e.g. 1000 samples) or when the visualization is not required to be real-time.

III. Proposed Method

The proposed framework combines both local and global XAI strategies for PD. Local XAI aims to explain individual

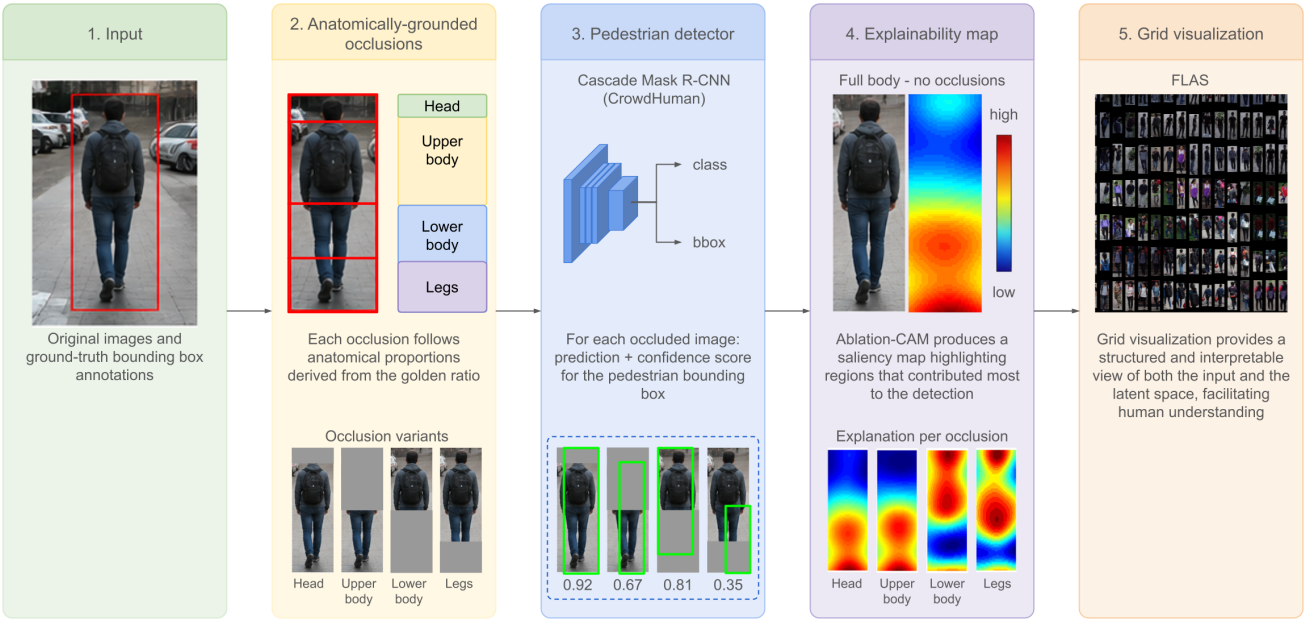


Fig. 2: Overview of the proposed explainability framework. 1-2) Pedestrian images are partitioned into anatomically meaningful regions based on golden-ratio-inspired body proportions. Region-wise occlusions are applied to the input images, 3) which are then processed by the pedestrian detector. 4) Ablation-CAM is used to generate saliency maps, 5) while grid-based visualization provides a structured global representation of the detector behavior under different occlusion settings.

detector decisions through semantically meaningful perturbations and saliency analysis, while global XAI aims to provide an overall understanding of the detector behavior across large sets of pedestrian samples.

As illustrated in Fig. 2, starting from input images with ground-truth (GT) bounding boxes (bboxes), we firstly apply semantically grounded occlusions by partitioning each pedestrian silhouette according to anatomical proportions derived from the golden ration (see Section III-A). Each occluded variant is then processed by a pedestrian detector, which produces predictions and confidence scores for the pedestrian instances. To interpret the model's behavior, we compute CAMs for each occlusion setting, highlighting the regions that contributed most to detection. Finally, we employ a grid-based visualization to organize input samples, enabling a structured and interpretable view of the learned feature space.

A. Anatomically-Grounded Occlusions

The golden ratio, is a mathematical proportion historically associated with harmony and aesthetic balance. In the context of human body measurements, it has often been used to describe idealized relationships between different anatomical segments, such as the ratio of total height to the position of the navel, or the proportions between limbs and torso. Classical representations, such as Leonardo da Vinci's Vitruvian Man (see Fig. 1), reflect this search for geometric order in the human form, suggesting that the body can be interpreted through consistent proportional rules [1].

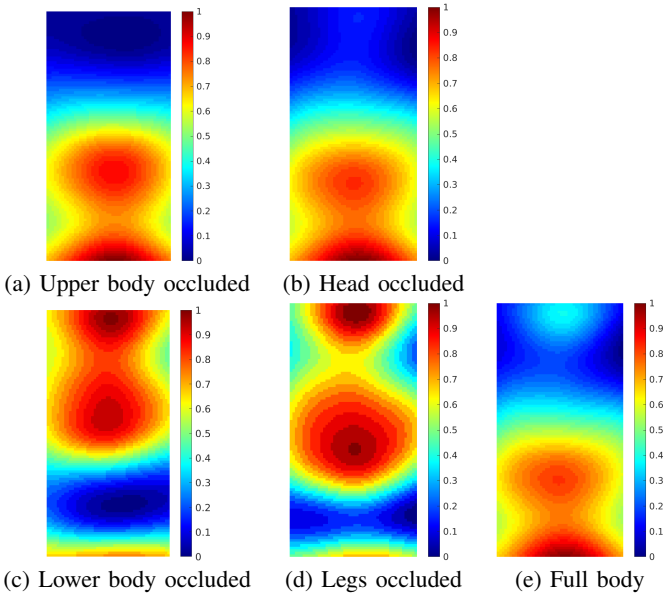
In this work, we focus on the body proportions relevant to the proposed occlusion experiments, namely: the head height corresponding to approximately $1/8$ of the body height, the upper body extending from the head to the torso region, the lower body corresponding approximately to the lower half of the body, and the leg region – from below the foot to below the knee – representing approximately $1/4$ of the body height. These proportions are used to define semantically meaningful occlusion regions for XAI analysis.

B. Local Explainability

Traditional XAI methods for CV generally rely on saliency maps or activation maps to highlight the regions used by the model during inference. However, while such methods indicate where the model focuses its attention, they often provide limited semantic interpretation regarding the role played by different human body regions in the final decision [8].

To address this limitation, we propose a semantically grounded local XAI strategy based on anatomically meaningful occlusions combined with Ablation-CAM visualizations. For each pedestrian GT bbox, four occlusion settings are generated according to the anatomical regions described in Sec. III-A: *head*, *upper body*, *lower body*, and *legs*. In addition, the original non-occluded bboxes are also considered for comparison purposes. Each occluded image is processed independently by the pedestrian detector, producing a detection confidence score and a predicted bbox.

To further interpret the detector behavior, we compute Ablation-CAM for each occlusion setting. The resulting activation maps highlight the image regions contributing most to PD under each anatomical occlusion setting. By comparing them



(a) Upper body occluded (b) Head occluded

(c) Lower body occluded (d) Legs occluded (e) Full body

Fig. 3: Average Ablation-CAM activation maps. The heatmaps highlight how the detector redistributes its attention depending on the visible body regions. Interestingly, although the full-body setting shows relatively limited attention on the upper body (e), these regions become significantly more important when lower body parts are occluded (c-d), suggesting the ability of the detector to rely on complementary semantic body cues when necessary.

with the corresponding detector average precision, it becomes possible to analyze how the detector redistributes its attention when specific semantic body regions are removed.

C. Global Explainability

The visualization of input data or the latent space of a model can be used to get a global understanding of the model's behavior. This is commonly done by projecting them in a 2d dimensional space and representing them with a scatter-plot. As previously presented, we favor to use grid-based visualization techniques to get a better representation of the data density and to use the screen space more efficiently. We can assume that close pedestrian bounding-boxes in the grid have a close feature representation. By visually analyzing the grid, we can get insights on the model's behavior, such as the presence of clusters of similar bboxes, the distribution of true positives and false positives, etc.

It is interesting to analyze the bboxes based on the anatomical properties of the pedestrians. Thus, according to the golden ratio described in Sec. III-A, we applied occlusions on the pedestrian GT bboxes to obtain the following test sets: *full body, lower body occlusion, upper body occlusion, head occlusion, leg occlusion*.

As such the grids are generated for each test set. A bbox can be used to extract the image patch corresponding to the bbox as well as the Ablation-CAM representation of the pedestrian detector (see II-A for more details on the Ablation-CAM representation). Both approaches are experimented.

IV. Experimental Protocol

A. Dataset

The experiments rely on the *P-DESTRE* dataset [16], a fully annotated UAV-based pedestrian dataset designed for PD, tracking, re-identification, and search tasks. The dataset was acquired using DJI Phantom four drones flying over outdoor university campus environments under realistic surveillance conditions, with varying viewpoints, camera pitch angles, illumination conditions, and pedestrian densities.

P-DESTRE contains video sequences recorded at 30 fps with a spatial resolution of 3840×2160 pixels. The dataset includes annotations for pedestrian bboxes at frame level, consistent pedestrian identities across acquisition sessions, head pose information, and multiple soft biometric attributes. In total, the dataset contains more than 250 identities observed across multiple sessions and viewpoints.

B. Pedestrian Detection

Pedestron [13] is a MMdetection based repository (MMDetection[1] is an open source object detection toolkit based on PyTorch and part of the OpenMMLab project[2]), that focuses on the advancement of research on PD. It provides a list of detectors, both general purpose and pedestrian specific, to train and test. It also provides pretrained models and benchmarking of several detectors on different PD datasets. The pretrained models are made available on the Pedestron website[3].

We assess pedestrian-detector XAI under synthetic occlusion using a Pedestron/MMDetection inference pipeline with Ablation-CAM. Pedestron provides different models for PD, the selected model is the Cascade Mask R-CNN, trained on the general human/person detection dataset: Crowd Human [24].

Given an input RGB frame I, the image is resized to 960×540 pixel. GT pedestrian annotations coordinates are mapped to the image. Let a pedestrian annotation be represented as $b = (x, y, w, h)$, where (x, y) is the top-left corner and (w, h) are width and height. For the head-occlusion setting used in the script, we define:

$$\Omega_{\text{head}} = [x, x + w) \times [y, y + \lfloor h/8 \rfloor).$$

The occluded image $\tilde{I}_{head}$ is produced by replacing Ω_{head} pixels with their local mean intensity (channel-wise). The experiments are repeated for occlusion variants: upper body (top half), lower body (bottom half), and legs (bottom-quarter) regions to obtain ($\tilde{I}_{upper}$, $\tilde{I}_{lower}$,$\tilde{I}_{leg}$).

GT pedestrian bboxes are used to define the anatomical occlusion regions rather than predicted detections. This choice ensures spatially consistent body partitioning across all experiments and avoids introducing additional variability caused by localization errors or missed detections. Using predicted

[1]Read more at https://github.com/open-mmlab/mmdetection
[2]Read more at https://openmmlab.com/
[3]Pedestron: https://github.com/hasanirtiza/Pedestron

bboxes could lead to inconsistent anatomical masking and would impact the analysis of the semantic contribution of different body regions.

The occluded image $\tilde{I}_\bullet$ is processed by the pedestrian detector. Predicted bboxes are represented by $b = (x, y, w, h, s)$, where s is the prediction confidence. Predicted bboxes are filtered by confidence:

$$\mathcal{B} = \{b_i \mid s_i > 0.3\}$$

where s_i is the detector score for box b_i. XAI maps are computed with Ablation-CAM. This yields a grayscale activation map M, which is resized to 960×540 pixel.

C. Local Explainability with Ablation-CAM

Local XAI experiments are conducted using the `pytorch-grad-cam` repository[4], which provides CAM-based XAI methods for CNN-based architectures. In this work, the Ablation-CAM [9] implementation is adapted to the Pedestron/MMDetection PD pipeline.

Following the implementation of the `pytorch-grad-cam` framework, feature-map importance is estimated by progressively ablating convolutional channels and measuring the corresponding decrease in the detector score. The method is configured using the `AblationLayerPedestron` ablation layer and the `pedestron_ablation_reshape_transform` reshape transformation provided by the modified implementation. The ratio of ablated channels is set to 1.0.

The resulting grayscale activation maps are resized to the original input image resolution (960×540). Average activation maps are computed independently for each occlusion configuration by averaging the generated Ablation-CAM maps over all pedestrian samples. Before averaging, each pedestrian is cropped using the GT bbox annotations and resized to a common spatial resolution of 50×100 pixels. This normalization step ensures spatial alignment between pedestrian samples and enables the analysis to focus on the contribution of coarse anatomical regions rather than fine-grained pixel-level variations.

In addition, pedestrian bboxes with an aspect ratio lower than 0.6 are excluded from the averaging process, as they typically correspond to partially visible pedestrians that do not contain complete body information. Including such samples would introduce structural inconsistencies in the averaged activation maps and reduce the interpretability of the resulting saliency distributions.

Fig. 3 presents the resulting average saliency maps for the different occlusion settings.

D. Global Explainability with FLAS

We have implemented all methods mentioned in the related works section, and selected the FLAS algorithm both for the overall quality of the visualization and for its computational efficiency. Several parameters are requested by the algorithm: the initial filter radius, the radius reduction factor, and the number of swap candidates; they are optimized with gridsearch on a subset of the dataset.

Euclidean distance is used to compare the feature vectors of the bboxes. The feature vector is build by concatenating CLIP features [21] and the color-histogram (this has empirically shown the best results among various tries including HOG features, raw RGB values, color-histogram alone or CLIP features alone). The pedestrian bounding is depicted using the raw pixel values or filtered version that keeps only the pixels within the 30% most important according to the heatmap.

The aspect ratio of the bboxes is preserved when displayed to a grid. We have chosen to generate grids that are visually squared (e.g. there are more images horizontally than vertically because of the aspect ratio of a pedestrian).

[4]Ablation-CAM: https://github.com/jacobgil/pytorch-grad-cam

Method	AP	Recall	Precision	F1 Score
Full body - orig. size	85.68	57.94	88.79	0.701
Full body - resized 1/4	86.51	57.38	90.28	0.701
Lower body occluded	68.37	49.74	80.52	0.615
Upper body occluded	35.25	30.22	59.96	0.401
Head occluded	81.98	55.09	87.45	0.674
Legs occluded	79.83	54.53	86.26	0.668

TABLE I: Pedestrian detection performance under different anatomically grounded occlusion settings. The results report Average Precision (AP), Recall, Precision, and F1-score, enabling the analysis of the contribution of different body regions to the detector performance.

V. Results and Discussion

To quantitatively evaluate the impact of the proposed anatomical occlusions on PD performance, we use the Average Precision (AP) metric, which is one of the standard evaluation measures for object detection. AP summarizes the precision-recall curve by computing the average detection precision over all recall levels. In practice, a predicted bbox is considered correct when its Intersection over Union (IoU) with a GT pedestrian annotation exceeds a predefined threshold. AP therefore jointly measures the ability of the detector to correctly localize pedestrians while minimizing false detections.

For each occlusion configuration (*full body*, *head occlusion*, *upper body occlusion*, *lower body occlusion*, and *leg occlusion*), the corresponding modified dataset is processed independently by the Cascade Mask R-CNN detector. Detector outputs are filtered using a confidence threshold of 0.3, and the resulting predictions are compared against the original P-DESTRE GT annotations. Precision, recall, F1-score, and Average Precision are then computed for each occlusion setting, enabling the quantitative analysis of the contribution of different anatomical regions to PD performance. The results are summarized in Tab. I.

Although the activation maps in Fig. 3 indicate that the detector focuses primarily on the lower body region, and in particular on the legs, the performance degradation is most pronounced when the upper body is occluded. This apparent contradiction highlights the complementary roles of different anatomical regions in the detection process. The

lower body provides strong local and discriminative cues, which explains the concentration of activation in this area. However, the upper body contributes essential global structure and contextual information that stabilizes the detection. When the upper body is removed, the model loses important spatial and geometric consistency, leading to a significant drop in performance. This suggests that PD relies not only on localized salient features but also on the overall coherence of the human body. This behaviour suggests that current XAI methods may overemphasize highly activated regions, while underestimating the importance of less salient but structurally critical regions.

Regarding global XAI, the resulting grid visualizations reveal several global organization patterns in the detector representation space. Large-scale structures are strongly influenced by illumination, contrast, and overall pedestrian silhouette appearance, indicating that global visual characteristics significantly contribute to the learned representation. In addition, neighboring regions in the grid frequently contain pedestrians with similar body visibility, posture, and viewpoint, suggesting that the detector also organizes samples according to semantically meaningful pedestrian properties. Transition regions between major clusters often correspond to visually ambiguous or partially occluded pedestrians, which may indicate more challenging detection conditions.

The thresholded grid visualizations provide additional insights into the global behavior of the pedestrian detector by emphasizing only the most discriminative regions identified by Ablation-CAM. Compared to the original RGB grids, the thresholded representations reduce the influence of background appearance, clothing color, and illumination conditions, allowing the organization of the feature space to reflect primarily the detector attention patterns. The resulting grids reveal that pedestrians are mainly grouped according to structural and anatomical cues, such as body silhouette, pose, and the spatial distribution of activated regions. In particular, strong activations are frequently concentrated around the lower body and leg regions, confirming the observations obtained from the local explainability analysis. Furthermore, the thresholded grids facilitate the identification of outliers and atypical activation patterns, corresponding to heavily occluded, truncated, or low-confidence pedestrian samples.

A zoomed view of the FLAS grids, is presented in Fig. 4 for readability purposes, while the complete high-resolution FLAS visualizations, including thresholded grids, as well as the t-SNE grid visualizations for comparison, are provided as supplementary material at: Supplementary Material[5].

Overall, the proposed grid-based visualization provides an interpretable global overview of the detector behavior and complements the local explanations obtained through Ablation-CAM.

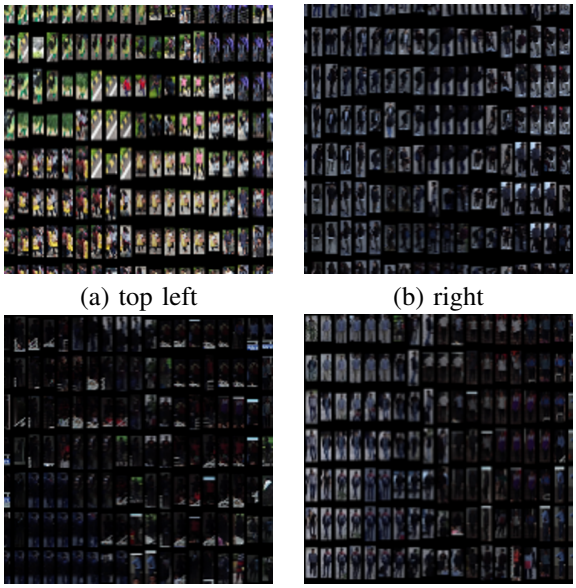
VI. Conclusion

The proposed approach addresses an important limitation of traditional XAI methods, which often highlight relevant image regions without providing explanations that are easily interpretable from a human perspective. By constraining perturbations according to meaningful anatomical body regions, the proposed framework enables a more intuitive understanding of the detector behavior.

Local explainability was achieved through anatomically meaningful occlusions inspired by Leonardo da Vinci's golden ratio and analyzed using Ablation-CAM saliency maps, while global explainability was investigated through grid-based visualization of pedestrian representations.

The experimental results showed that different anatomical regions contribute differently to pedestrian detection performance. Although the Ablation-CAM analysis revealed that the detector attention is strongly concentrated around the leg region, the largest performance degradation was observed when the upper body was occluded. This suggests that pedestrian detectors rely not only on the most visually salient regions, but also on complementary semantic body structures required to preserve the global consistency of the human silhouette.

The proposed global explainability analysis further demonstrated that grid-based visualizations can reveal meaningful organization patterns in the detector representation space, including similarities related to body structure, visibility, and activation distributions. In particular, thresholded CAM-based grids were shown to reduce appearance bias and provide a more faithful representation of the detector attention mechanisms.



(a) top left (b) right (c) bottom (d) transition

Fig. 4: FLAS-based global explainability: (a) top-left region contains brighter images with higher contrast and strong sunlight; (b) right region is dominated by dark clothing and lower illumination; (c) bottom region contains more back-facing pedestrians; (d) transition regions between clusters contain more ambiguous or partially occluded pedestrians. Complete HR available at Supplementary Material[5].

[5]Grid visualizations: https://galdi.eurecom.io/ExFMA2026_supplementary_material.html

References

[1] Evon Abu-Taieh and Hamed S. Al-Bdour. A human body mathematical model biometric using golden ratio: A new algorithm. In Jucheng Yang, Dong Sun Park, Sook Yoon, Yarui Chen, and Chuanlei Zhang, editors, *Machine Learning and Biometrics*, chapter 7. IntechOpen, London, 2018.

[2] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[3] Kazi Ahmed Asif Fuad, Pierre-Etienne Martin, Romain Giot, Romain Bourqui, Jenny Benois-Pineau, and Akka Zemmari. Features Understanding in 3D CNNs for Actions Recognition in Video. In *Tenth International Conference on Image Processing Theory, Tools and Applications, IPTA 2020*, Paris, France, Oct. 2020.

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, 2015.

[5] Kai Uwe Barthel, Florian Tim Barthel, and Peter Eisert. Permutation learning with only n parameters: From softsort to self-organizing gaussians. In *2025 33rd European Signal Processing Conference (EUSIPCO)*, pages 1892–1896, 2025.

[6] Kai Uwe Barthel, Florian Tim Barthel, Peter Eisert, Nico Hezel, and Konstantin Schall. Creating sorted grid layouts with gradient-based optimization. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1199–1206, 2024.

[7] Kai Uwe Barthel, Nico Hezel, Klaus Jung, and Konstantin Schall. Improved evaluation and generation of grid layouts using distance preservation quality and linear assignment sorting. In *Computer graphics forum*, volume 42, pages 261–276. Wiley Online Library, 2023.

[8] Nuo Chen, Ming Jiang, and Qi Zhao. Explainable saliency: Articulating reasoning with contextual prioritization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9601–9610, 2025.

[9] Saurabh Desai and Harish G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 972–980, 2020.

[10] Ohad Fried, Stephen DiVerdi, Maciej Halber, Elena Sizikova, and Adam Finkelstein. Isomatch: Creating informative grid layouts. In *Computer graphics forum*, volume 34, pages 155–166. Wiley Online Library, 2015.

[11] David Gunning. Darpa's explainable artificial intelligence (xai) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page ii, New York, NY, USA, 2019. Association for Computing Machinery.

[12] Adrien Halnaut, Romain Giot, Romain Bourqui, and David Auber. Vrgrid: Efficient transformation of 2d data into pixel grid layout. In *2022 26th International Conference Information Visualisation (IV)*, pages 11–20. IEEE, 2022.

[13] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11328–11337, June 2021.

[14] John Healy and Leland McInnes. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4(1):82, 2024.

[15] K. S. Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C. King, and Kevin W. Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020.

[16] S. V. Aruna Kumar, Ehsan Yaghoubi, Abhijit Das, B. S. Harish, and Hugo Proença. The p-destre: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *Trans. Info. For. Sec.*, 16:1696–1708, Jan. 2021.

[17] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, and M. Angelini. State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum*, 42(1):319–355, 2023.

[18] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[19] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[20] Yangyang Qu, Massimiliano Todisco, Chiara Galdi, and Nicholas Evans. Fair-gate: Fairness-aware interpretable risk gating for sex-fair voice biometrics. In *2026 14th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2026.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[23] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[24] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

[25] Grant Strong and Minglun Gong. Self-sorting map: An efficient algorithm for presenting multimedia data in structured layouts. *IEEE Transactions on Multimedia*, 16(4):1045–1058, 2014.

[26] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[27] Martin Winter, Werner Bailer, and Georg Thallinger. Demystifying face-recognition with locally interpretable boosted features (libf). In *2022 10th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2022.

[28] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.