OPTIMIZING PEDESTRIAN ATTRIBUTE RECOGNITION IN LOW LIGHT CONDITIONS

Amr Alomari, Charles Cheung, Jonathan Sin
Department of Computer Science, University of Toronto
{amr.alomari, charles.cheung jonathan.sin}@mail.utoronto.ca

ABSTRACT

Pedestrian Attribute Recognition (PAR) models trained on visible-spectrum datasets perform poorly in low-light and infrared settings due to severe domain shift in illumination, contrast, and crop geometry. To address this, we build a modular VIS–IR pipeline that pairs a cross-modal detector (DEYOLO) with multiple PAR models trained on PA-100K (PP-LCNet, ResNet-18, VTB). We analyze the structural differences between PA-100K and LLVIP and show how luminance, aspect ratio, and low-light degradation directly affect attribute prediction. Enhancement-based domain mapping (log-normalization, VIIS fusion) is explored but rejected due to distortion of semantic details critical for PAR. We then evaluate the three models on a small set of manually curated LLVIP crops and introduce two lightweight ensemble strategies: a majority-vote system and a category-specialized expert router. On this preliminary sample, both ensembles show improved performance over individual models, with the expert approach achieving the strongest gains despite zero retraining. These early results suggest that even without full domain adaptation, a modular pipeline plus targeted ensembles can help recover some low-light PAR performance, pointing toward practical applications in retail security, home surveillance, and incident review.

1 Introduction

The demand for surveillance, detection and attribution systems has increased, particularly surrounding pedestrians on the street, in low-light conditions, and applying attribution. In the modern day, there is an ever-increasing need for intelligent surveillance and peacekeeping and research into psychological and social behaviors. However, tight-knit integration of such systems is a problem that still allows for further refinement.

Additionally, existing solutions seldom focus on low-light conditions, which can be problematic for use cases at nighttime due to loss of effective target areas [1]. Surveillance needs frequently rise during nighttime because darkness provides cover, leading to an increased risk of specific crimes [2]. This problem is not adequately addressed with PAR-only solutions.

Difficulty can arise when Pedestrian Attribute Recognition (PAR) systems that are built for recognition tasks such as clothing determination (as needed for social behavioral studies or profiling systems) simply do not have additional information required for the integration needs of surveillance systems to include other varied attributes (such as whether or not some pedestrians might be carrying weapons).

To alleviate the problem, we introduce a modular, custom pipeline that integrates several high-performance models through an accessible, minimal and user-friendly workflow that allows a user to easily execute this task. At each step, modularity allows for an easy replacement for any section of the pipeline. We demonstrate the pipeline with DEYOLO [3] for detection, and an ensemble of PAR models (PP-LCNet [4], VTB [5] and Resnet-18 [6]) trained on different attribution types (clothing, age, etc).

1.1 Motivation

Our goal is to improve Cross-Domain Pedestrian Attribute Recognition, specifically in low-light conditions. However, another goal is to increase accessibility and ease of usage for proposed solutions to this type of problem; by creating

a full-flow Detection-and-PAR pipeline using LLVIP and PA100K-Trained models [1, 7] that can run on consumer hardware with relatively low compute. A preset combination of models with great low-light performance shows the modular nature of the pipeline, and shows users how to modify the pipeline to suit their needs.

There is suitable demand for an accessible solution to this problem. We acknowledge that not every person is capable of programmatically automating their workflows; our intended usage is for a user to upload an image(-pair) to the pipeline, wait for processing, then see bounding boxes around pedestrians and their attributes. In addition, several use cases show the need for such solutions:

- Businesses:
  - Shoplifting: Stores can add a shoplifting detection PAR model to the pipeline in order to identify shoplifter attributes in dimly lit areas. Notable for exterior cameras at nighttime.
  - Customer Profiling: Automatic aggregated (and deidentified) data-gathering for gender, age, and clothing demographic of customers.
  - Store Layout: Enable comparisons of attribute distributions between different entrances, aisles, or store areas (e.g., "more formal clothing near the officewear section").
- Personal:
  - Home Security: Intruder identification and fast on-site attribution and description generation. Can be combined with other security solutions such as doorbell cameras or home integration.
- Law Enforcement:
  - Missing Persons Search: Attribute detection allows for better searches of missing people
  - Incident Review: For a particular time window, quickly filter cameras by attributes to reduce manual scanning.

2 Data

2.1 LLVIP

The LLVIP dataset [1] is an annotated paired-image dataset. It consists of 15488 pairs of Visible-Spectrum and Infrared photos taken by the same camera in strictly the same time and space, and approximately the same photographic settings (eg. FOV). This is possible due to a Dual-Spectrum camera with two sets of lenses integrated into the camera, as well as image post-processing and registration, to ensure the images and their features line up correctly.

The content of the dataset is a majority of nighttime street images taken overhead of pedestrians. Pedestrians in the dataset are labeled, and they are the only existing class label. Training and Validation sets are given existing as separate directories in the dataset. For an example of a data point, please see Figure 1.

(a) visible image with annotations (b) infrared image with annotations

Figure 1: LLVIP Data Example

2.2 PA-100K

The PA-100K dataset [7] is one of the largest and most commonly used pedestrian attribute recognition benchmarks. It contains 100,000 street-scene images collected from 598 distinct surveillance cameras. Each image is annotated with 26 binary pedestrian attributes we have split into five categories: *gender*, *age*, *orientation*, *upper-body clothing*, *lower-body clothing*, and *accessories*.

Unlike paired datasets, PA-100K includes heterogeneous poses, backgrounds, illumination conditions, and occlusion levels, making it a challenging benchmark for evaluating model robustness in real-world scenarios. The dataset is pre-split into training, validation, and test sets (80K / 10K / 10K), and follows a standardized evaluation protocol widely adopted in PAR research.

In contrast to LLVIP's dual-modality (VIS–IR) paired images, PA-100K provides only single-modality visible-spectrum images. This lack of a corresponding infrared view increases the difficulty of cross-domain generalization, since attribute cues that may be more salient in IR (e.g., outlines, heat-based contrast, nighttime visibility) are unavailable. As a result, applying PA-100K–trained models to IR or low-light datasets requires additional mapping, and adaptation techniques.

2.3 PaddleX Custom Dataset(same classes as PA-100K)

The PaddleX custom Dataset is a dataset with labels with the same attributes as PA-100K, which is useful for model comparison and testing. Given that this dataset is closed source, not many assumptions can be made about it. In this paper we compare the model trained on this dataset's performance to the others trained on the open PA-100K dataset.

3 Data Analysis and Cleaning

3.1 Overview and Challenges

The PA-100K dataset provides PAR labels. However, the LLVIP dataset only provides class labels for bounding boxes and does not provide attribute labels. To mitigate this, we selected a sample of 10 random images from the cropped LLVIP set given by DEYOLO and manually verified labels to compare them to the models' predictions.

3.2 Analysis

In order to understand the performance gaps observed between PA-100K–trained models and their outputs on LLVIP, we conducted a systematic analysis of the underlying visual differences between the two datasets. A preliminary inspection revealed that LLVIP images differ substantially in both illumination and pedestrian crop geometry. Specifically, LLVIP contains predominantly nighttime or low-light imagery, while PA-100K features mostly daytime visible-spectrum scenes. In addition, the DEYOLO-generated LLVIP crops show far greater variation in bounding-box size and aspect ratio compared to PA-100K.

To quantify these differences, we computed luminance for each crop using the standard formula $\text{lum} = 0.299R + 0.587G + 0.114B$, and measured aspect ratio as the ratio of height to width. These distributions provide insight into the domain shift faced by PA-100K–trained models when applied to LLVIP, and help contextualize the models' attribute-prediction behavior under differing visibility and geometric conditions.

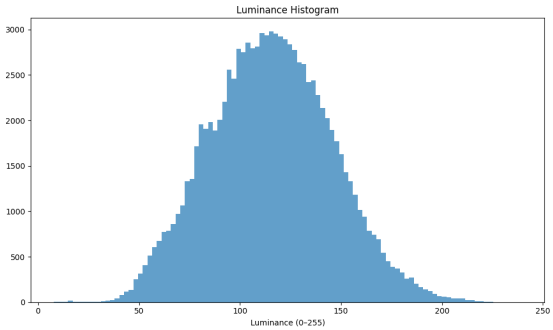

Figure 2: *
PA-100K luminance distribution

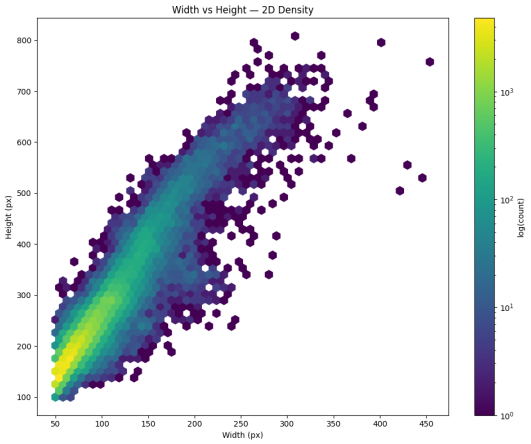

Figure 3: *
PA-100K aspect ratio distribution

Figure 4: Luminance and aspect ratio statistics for PA-100K.

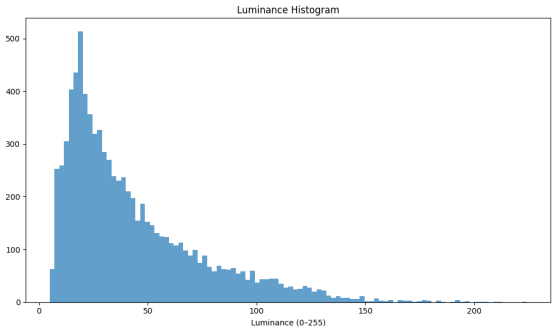

Figure 5: *
LLVIP-DEYOLO luminance distribution

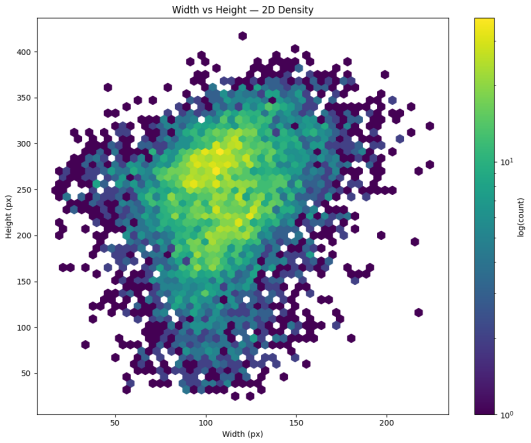

Figure 6: *
LLVIP-DEYOLO crops aspect ratio distribution

Figure 7: Luminance and aspect ratio statistics for LLVIP DEYOLO crops.

We then bounded our output dataset to two bounding lines used in our evaluation are:

$$Upper = 2.3333\,w + 180,$$

$$Lower = 3.1538\,w - 157.6923.$$

which cut the total cropped validation set by 32% from 8302 images to 5639.

These lines were chosen to closer model the distribution of heights and widths in the PA-100K dataset.

3.3 Selection

By restricting the test set, we were then able to go in and manually select 10 images that showed full body crops and were not extremely dark, which would allow the models to evaluate on all attributes in their output set in a best case scenario.

Figure 8: Example of Evaluation Image

3.4 Other Mapping Methods Explored

Log-domain luminance normalization

Before evaluating model performance on LLVIP, we applied a two–stage data cleaning process to remove invalid crops and normalize illumination. These steps were used solely to standardize the LLVIP inputs; no model was retrained on the processed data.

LLVIP images are substantially darker than PA-100K because they are captured in nighttime conditions. To reduce extreme illumination differences during testing, we applied a simple luminance normalization step. Using luminance samples from both datasets, we computed the mean and standard deviation of their log-luminance distributions. Each LLVIP luminance value Y was then mapped to the PA-100K range using:

$$\log Y' = \mu_{\text{PA}} + \sigma_{\text{PA}} \left(\frac{\log Y - \mu_{\text{LLVIP}}}{\sigma_{\text{LLVIP}}} \right), \qquad Y' = \exp(\log Y').$$

Figure 9: Example of Log Transformation on Figure 8

This method, however, took away from the contrast which must also be evaluated and adjusted in order to allow the models to perform accurately compared to the original image

VIIS: Visible and Infrared Information Synthesis for Severe Low-light Image Enhancement

We also explored a VIIS method [8], which synthesizes enhanced images by fusing visible-spectrum and infrared cues through a latent-diffusion framework. The approach is designed to recover structure, illumination, and semantic detail under severe low-light conditions by learning a joint VIS–IR representation and generating a brighter, more informative output. In prior work, VIIS demonstrates strong enhancement capability, producing visually sharp reconstructions that outperform classical low-light enhancement techniques.

Despite these advantages, VIIS proved unsuitable for our evaluation pipeline. First, the model operates on fixed-size 512×512 inputs, requiring aggressive resizing, which could only be dealt with by either moving the enhancement before the DEYOLO processing stage and replacing the VIS image in the crops with the corresponding output VIIS image (Severely reduces the high quality of LLVIP images), or by adding padding around the pedestrian crops (which

greatly reduces the quality of the output image). Furthermore, on our tested images, the VIIS model substantially alters the geometry and coloring of the pedestrians, which is incompatible with our attribute prediction task that rely on fine-grained visual detail. The diffusion-based synthesis introduces a "painted" appearance and modifies edges and boundaries, occasionally altering clothing contours and object outlines. Because our goal is to assess model performance on the *original* image content, such generative alterations would invalidate any attribute predictions made on the enhanced images. Suggested solutions to this method would require retraining the model to better preserve geometry and coloring for images similar to the LLVIP dataset, making it work with dynamically sized vertical image crops, as well as either upscaling, or maintaining the high resolution that LLVIP provides.

For these reasons, we opted not to evaluate on either enhancement method.

Figure 10: Example of a VIIS-Altered Image

4 Methodology

The pipeline is built around modularity concerning the two required stages ("Detection" and "Pedestrian Attribute Recognition (PAR)"). For the proceeding passage, please consider the stages in context of the preset models chosen for the pipeline implementation. These are:

- (detection) DEYOLO [3], which is a supervised object detection model. It allows multiple inputs on YOLO using special attention modules (DECA and DEPA) as well as a bi-directional decoupled focus module for cross-modality enhancement.

  Historically, multi-spectrum detection models have attempted a workflow that fuses input images before applying detection. DEYOLO uses special attention mechanisms to directly look at both inputs, making full use of both spectra instead of the conventional fusing method. Great in low-light conditions, and provides accurate crops for inputs going into the next PAR stage.
- (PAR) Ensemble of models: This allows different PAR models to specialize in different attribute types, such as clothing, age, etc. The results are derived as combinations of the outputs of the individual PAR models.
  - PP-LCNet [4], lightweight convolutional neural network designed for high efficiency on CPU devices. PP-LCNet uses depthwise separable convolutions, optimized channel configurations, and structural re-parameterization to achieve strong accuracy–latency tradeoffs. In our pipeline, PP-LCNet serves as a compact PAR model specialized for clothing-related attributes, offering fast inference while maintaining competitive performance.
  - VTB [5]: Uses a ViT-Base backbone as the visual encoder and a simple text embedding module for attribute descriptions, followed by a transformer that jointly reasons over visual tokens and attribute tokens. We use the official PA-100K implementation and checkpoint as a strong transformer-based PAR baseline. This is the most computationally expensive model that was trained, and serves as a good benchmark to compare the other, smaller models to.
  - ResNet-18 [6]: A compact CNN backbone with residual blocks and global average pooling. In our pipeline, ResNet-18 is used as a student model in a teacher–student configuration inspired by the SNN-PAR distillation framework: VTB serves as the teacher, and the ResNet-18 student is trained on PA-100K to match both the ground-truth attribute labels and VTB's soft logits. This yields a lighter CNN that partially inherits the transformer's behaviour while being cheaper to run at inference time.

The architecture consists of:

- **Input**: Visible-spectrum and Infrared image-pairs
- **Detection Stage**: A module which must provide bounding box predictions from the given input. The preset model is DEYOLO.
- **Bounding Box Predictions and Post-detection Processing**: The source image is cropped according to each prediction. Each resulting image is then passed indivudually into the next stage.
- **Pedestrian Attribute Recognition**: An ensemble of PAR models specialized in different attribute types. Different models (PP-LCNet, and VTB-to-ResNet-18 in teacher-student configuration) are ensembled for greater accuracy.
- **Output**: Attribute predictions per model, ensemble predictions, and processing times

Please refer to Figure 11 for the architecture diagram.

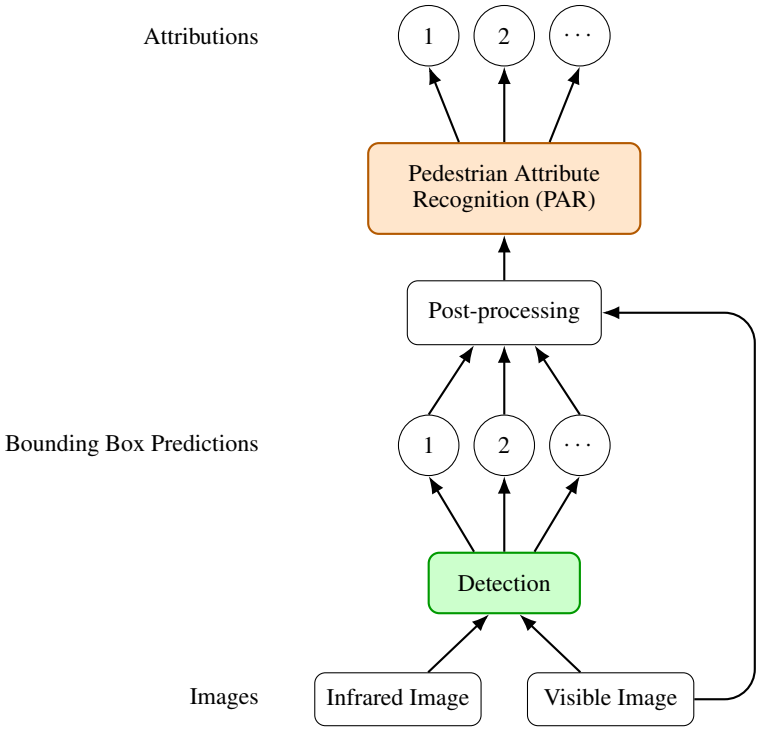

Figure 11: Pipeline architecture: modular stages as colored nodes.

4.1 Decisions

The modularity of the system allows for encapsulation and separation of concerns on the part of detection and attribution. Naturally, this allows users to swap in any preferred models, as well as for the capability to debug the system effectively during implementation. This provides accessibility and ease of modification to any needs.

Ensembling PAR models allows different PAR models to specialize in different attribute types, such as clothing, age, etc. The results are derived as combinations of the outputs of the individual PAR models. This allows for effectiveness in time constraints as the performance of many smaller PAR models can be comparable to simply training a large scale PAR model. This also allows for fast iteration and research and testing different potential PAR model presets.

After bounding box predictions are acquired, the decision was made to post-process the source image into cropped versions according to each prediction. This is because the chosen preset PAR models are trained on similar, zoomed-in shots of subjects, and therefore it is natural to provide similar such images as input to the PAR stage.

4.2 Novel aspects

There are many existing PAR pipelines integrating pedestrian detection which extract pedestrian attributes from surveillance cameras. Our novelty lies in three aspects:

1. Evaluating different combinations of detection models and PAR models, and comparing their performance with both our custom pipeline and existing pipelines;
2. Designing the pipeline to focus specifically on low-light environments, a scenario that most current PAR pipelines do not adequately address.
3. Incorporating an ensemble framework into the PAR stage to leverage the strengths of multiple models and improve overall accuracy.

4.3 Incompleteness

Given the short time frame for implementation, the sample testing and validation set for our ensemble models was extremely small (10 images) and the ensemble models were only pseudo-implementations to give us an idea of which method was more worth pursuing.

Given the data for the ensembles, more work would be need to be done to:

- Label a larger portion of LLVIP for PAR tasks (Depending how large the sample, this could also enable training on the sample) - Re-evaluate the ensemble model on this larger set - Clean datasets per model (PA-100K), to hyper-specialize in their respective categories, reducing compute and increasing accuracy in the respective task - More complex ensemble logic, such as taking into account the confidence scores to determine more robustly how much to trust a model in the given category

Furthermore, more research would need to be done specifically for the IR + VIS -> VIS fusion task which there is not much research in (Producing enhanced/day images from night IR + VIS images)

4.4 Link

Please find the implementation at the following GitHub repository:

`https://github.com/alphabot12340/LLPAR_pipeline`

Instructions are included in the top-level `README.md` file.

5 Evaluation and Analysis

One of the strengths of the modularity of the pipeline is that analysis and optimization can be parallelized across the various stages, as well as encapsulation so that no optimization of any stage prevents an ongoing optimization in another. However, due to the nature of the problem, this is also a weakness, as an ideal joint-loss configuration is not possible and the accuracy of the latter stages of the pipeline depend on the accuracy of the former stages, which results in a detrimental cascade effect. As such, we present our analysis in separate parts, in the context of our present models as discussed in Section 4.

5.1 Training and Inference

Training and inference of models was done in an approximate equal spread of responsibility for models. Assorted methods were used to ensure reliability and reproducibility of models and pipeline.

Training robustness was handled well by our decision to maintain reproducibility across environments (local, cloud) across our team. Training and inference worked well on both. However, given our consumer hardware and free-trial of Google Colab, the speed of training and inference was quite slow. But we were satisfied, given comparison to the technical debt and regression, and high program translation starting cost of dealing with academic and datacenter solutions such as the university computing block, or Compute Canada being run through terminal. We were able to iterate extremely quickly with compute being at our fingertips, which worked well with our fail-fast work ethic.

5.1.1 Detection Stage

Our team used local resources (RTX4080S) and cloud resources (Google Colab) to train the detection model (DEYOLO). The implementation used PyTorch, and as such, we made sure to use CUDA-enabled versions of the PyTorch package to accelerate training and inference. A small, randomly-selected subset of 1000 data points was further selected out of the training set due to cost constraints on cloud resources and local resources. The Nano-size DEYOLO model was used to respect time and cost constraints, with the used hyperparameters similar as suggested by original repository (`epochs: 100, imgsz: 640, patience: 50, batch: 16, seed:0, lr0: 0.01, lrf: 0.01, momentum: 0.937, weight_decay: 0.0005, warmup_epochs: 3.0, warmup_momentum: 0.8, warmup_bias_lr: 0.1, box: 7.5, cls: 0.5, dfl: 1.5`).

DEYOLO Training plots can be found at Figure 12.

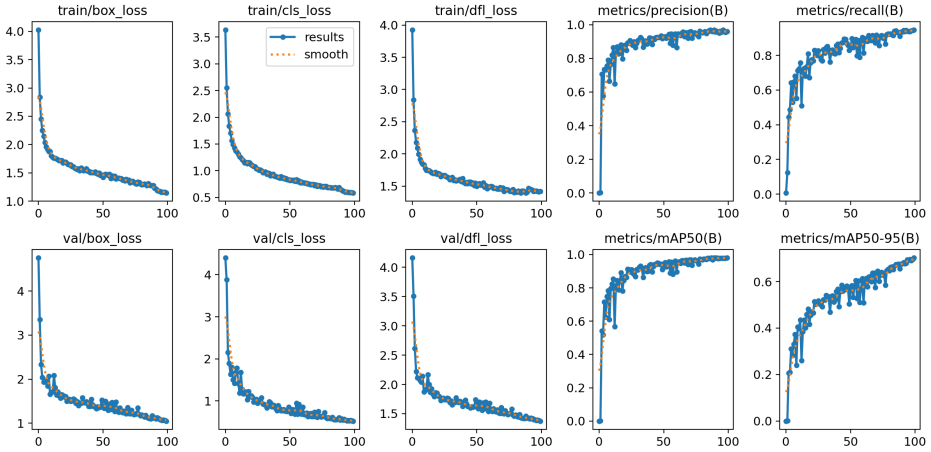

Figure 12: DEYOLO Training Plots

Inference was performed straightforwardly with `model.predict` syntax, at `confidence=0.25`.

5.1.2 PAR Stage

As LLVIP does not contain attribution labels, training against it is an unsupervised task and we are not privy to the same level of training analysis as the Detection stage.

5.2 Detection Analysis and Evaluation

Optimizing detection models was especially important because the output of the detection stage is directly the input of the later PAR stage. as such, a larger emphasis was placed on higher accuracy.

However, analysis of the detection stage can also suffer because there is only one existing class label. As such, analyses conventionally related to classification should be interpreted with reservation.

5.2.1 Overall Accuracy

The F1-Confidence Curve and Precision-Recall Curve for DEYOLO can be found at Figure 15.

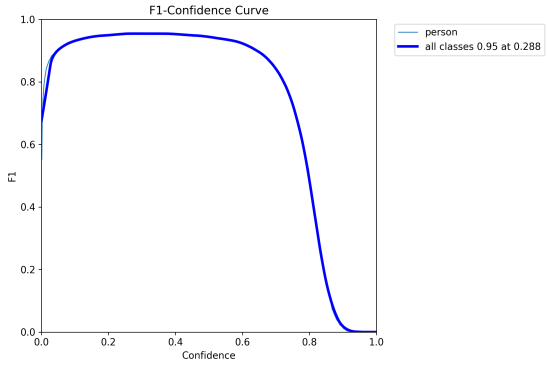

Figure 13: DEYOLO F1-Confidence Curve: Best performance at `confidence=0.288`

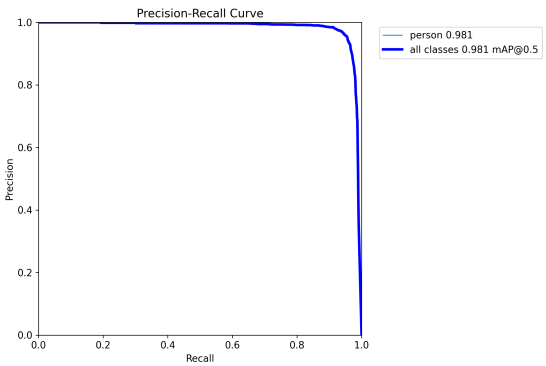

Figure 14: DEYOLO Precision-Recall Curve at `mAP@0.5`.

Figure 15: DEYOLO F1-Confidence Curve and Precision-Recall Curve.

The Confusion Matrix and Labels Plot for DEYOLO can be found at Figure 18.

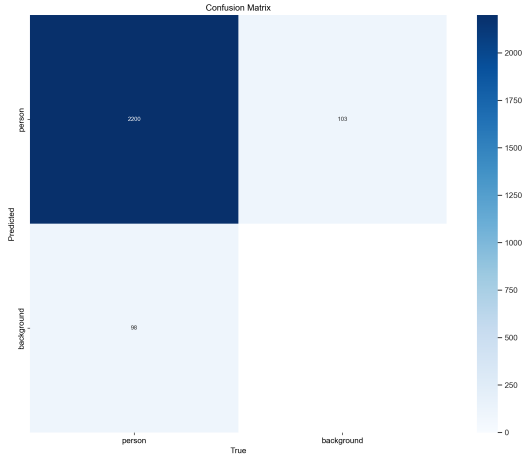

Figure 16: DEYOLO Confusion Matrix: single-class "person" performance against null class "background"

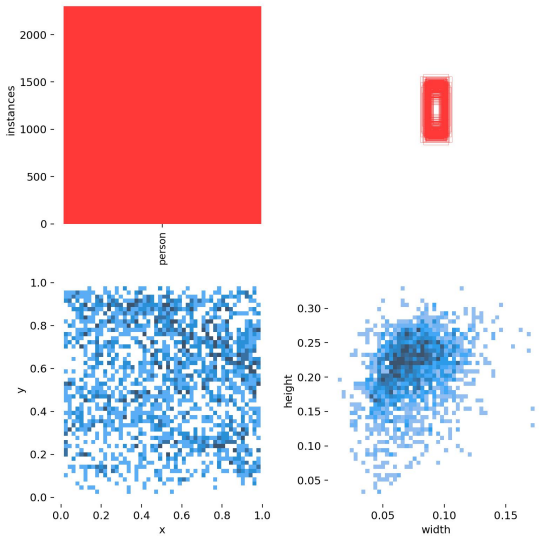

Figure 17: DEYOLO Labels Plot: Top-left showing single-class bar chart, top-right showing overlay of all bounding boxes predicted. Bottom-left showing distribution of bounding box predictions across normalized dimensions of input images, bottom-right showing height-width distribution of bounding box predictions.

Figure 18: DEYOLO Confusion Matrix and Labels Plot.

5.3 PAR Analysis and Evaluation

Evaluation was done on the PAR models and the data sample selected in order to determine accuracy of the models overall and on category subsets we have selected

Overall Accuracy

Model	Precision	Recall	F1 Score
VTB	0.772	0.786	**0.779**
PP-LCNet	0.673	0.700	**0.686**
Resnet-18	0.627	0.712	**0.667**

Figure 19: Evaluation of models

Category Splitting

In order to maximize PAR accuracy across the 26 output attributes given for the PA-100K dataset, we split the attributes into 6 categories:

Table 1: Attribute Categories Used for Model Routing

Category	**Attributes**
Personal	Age18–60, AgeLess18, AgeOver60, Female
Image / View	Back, Front, Side
Top Clothing	LongSleeve, ShortSleeve, LongCoat, UpperLogo, UpperPlaid, UpperSplice, UpperStride
Bottom Clothing	Trousers, Shorts, Skirt&Dress, LowerPattern, LowerStripe
Accessories	Backpack, Boots, Glasses, HandBag, Hat, ShoulderBag
Misc (ignored)	HoldObjectsInFront

The misc table is ignored because evaluation showed that none of the models could accurately predict this attribute. Further investigation is warranted, but for our paper, we choose to ignore this and focus on the rest.

The following is an evaluation on the three models we trained on the above categories:

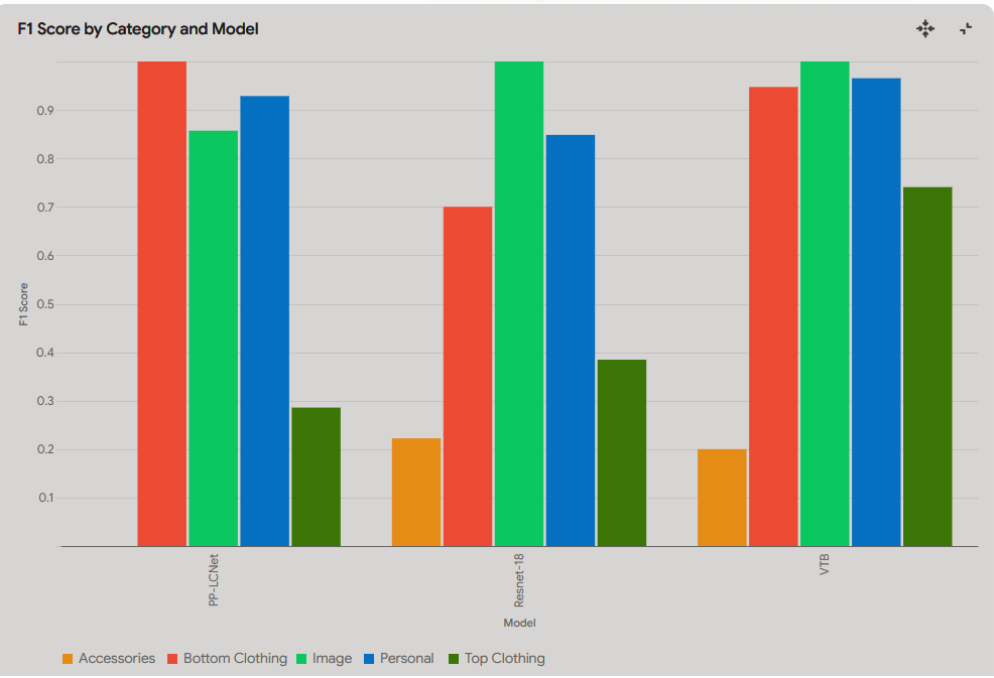

Figure 20: Model Evaluation Categories

Analysis of Category Performance

Given these metrics, the models performed fairly differently on each category. VTB was the strongest overall and consistently performed well on every attribute relative to the other models. RESNET and PP-LCNet models performed fairly similarly, however did have minor strengths, such as the Resnet model being able to predict accessories better than the PP-LCNet model.

Given these results 2 novel ways of combining the outputs were implemented:

Ensemble Voting

Given that the VTB was the strongest overall, PP-LCNet and Resnet were the weaker models with relative strengths, this ensemble model allows each model to vote on each attribute. If an attribute receives >= 2 votes, the ensemble will predict yes, else no.

This approach is generalizable to ensembles with strong and small, specialized models. The drawback is the addition of compute for the tradeoff of the potential for a relatively small gain of accuracy.

Ensemble Experts (MoE-like approach)

For each category prediction, the ensemble will trust the model that performed the best in the evaluation we conducted. The pro of this approach is that it decreases potential compute and will more greatly increase accuracy when hyper specialized models are run. Furthermore, the categorization allows the easy addition of attributes into our ensemble prediction model. The drawback is that the results shown may not apply when a greater sample size is taken.

Pseudo Implementation of Proposed Models

The models were trained on a python script with the above algorithms, and re-evaluated on the same sampled data to see how much accuracy could potentially increase.

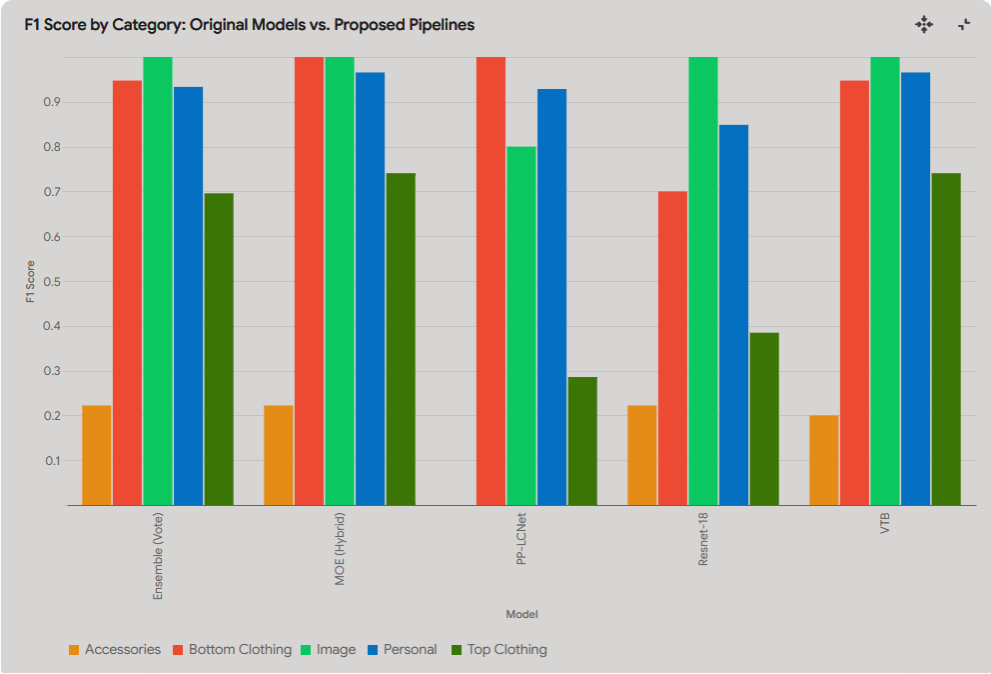

Figure 21: Proposed Model Evaluation

Analysis of Proposed Models Both models were able to outperform the original model, with the Ensemble expert model outperforming the MoE voting by a small margin. Given that Retraining each of these models on a smaller output set specialized to their respective categories would reduce compute, as well as potentially increase accuracy (as shown in this pseudo implementation), this is a path that may warrant further research.

6 Contribution

All: General research. Writing and presentation.

Jonathan: Pipeline design, setup and implementation. Setting up git repo, helping data processing of LLVIP dataset, research and help coding of DEYOLO model, implementing and researching PAR models for second phrases, implementing PP-LCNet model to the pipeline, PAR models testing and compare output data, testing DEYOLO model inference and training in colab, novelty research.

Amr: Planning, training, and implementing VTB and ResNet Models, input data analysis, data mapping exploration, VIIS application, evaluation of PAR models and analysis on the output, data aggregation and graphing output of PAR models, implementation of novel combination of par models, and evaluation of them.

Charles: Data processing LLVIP, researching preset detection model. Implementation, training, inference, evaluation, data analysis, and optimization of DEYOLO model. Pipeline design, post-processing logic, implementation of detection and post-processing pipeline stages.

7 Acknowledgements

Our project was supported by the Department of Computer Science through the CSC490 Capstone course.

References

[1] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021.

[2] Jim Uttley, Rosie Canwell, Jamie Smith, Sarah Falconer, Yichong Mao, and Steve Fotios. Does darkness increase the risk of certain types of crime? a registered report article. *PLOS ONE*, 20(6):e0324134, 2025.

[3] Yishuo Chen, Boran Wang, Xinyu Guo, Senbin Zhu, Jiasheng He, Xiaobin Liu, and Jing Yuan. Deyolo: Dual-feature-enhancement yolo for cross-modality object detection. In *International Conference on Pattern Recognition*, 2024.

[4] Cheng Cui, Tingquan Gao, Shengyu Wei, Yuning Du, Ruoyu Guo, Shuilong Dong, Bin Lu, Ying Zhou, Xueying Lv, Qiwen Liu, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. Pp-lcnet: A lightweight cpu convolutional neural network, 2021.

[5] Xinhua Cheng, Mengxi Jia, Qian Wang, and Jian Zhang. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[7] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis, 2017.

[8] Jing Zhang, Yuting Wang, Xuan Liu, et al. Viis: Visible and infrared information synthesis for severe low-light image enhancement. *arXiv preprint arXiv:2403.18916*, 2024.