# Multimodal Detection of Atomically-thin Semiconductor Materials

Jun Wang Jenny Hu Xueqi Chen
Stanford University
{junwang9, jingjinh, xqc1998}@stanford.edu

## Abstract

*We explored the advantage of detecting atomically-thin two-dimensional (2D) semiconductor material with two modalities: visible image photoluminescence image. We collected a set of images of exfoliated 2D materials, and created an annotated dataset, PLVIS2D. The single-modal baseline models for PLVIS2D are both outperformed by our multimodal models. The best model in our experiments adopted a cross-modality fusion transformer.*

## 1. Introduction

Object detection in computer vision has achieved amazing tasks and has broad application. A notable advantage of computer vision over human vision is the capability of processing multiple images of the same scene, especially when they are produced from multiple imaging modalities, such as thermal infrared (IR), x-ray, LiDAR, aside from imaging at visible wavelengths. Given multiple images of the same scene, humans have to inspect them sequentially, relying on short-term memory to correlate them and to further process, while these are not limiting factors for computers.

Coordinating information in images from multiple modalities have brought challenges to the field of computer vision. A general but core task for multimodal object detection is to create methods that fully utilize the complementariness of different modalities to achieve higher accuracy than with each single modality.

Another task is to generate a single but informative image from the multimodal sources, referred as cross-modality image fusion, a representative work for which is DenseFuse [7].

It is a practical topic for improving interpretability and for developing human-machine collaboration in a complex environment. For instance, image fusion of thermal-IR and optical images can significantly improve night-time surveillance quality without holding much more data. Meanwhile, cross-modality image fusion has inspired methods in multimodal object detection [10] .

Multimodal object detection has wide applications. For instance, studies combining thermal-IR imaging with the usual RGB camera vision have overcome challenges brought by low-light conditions for pedestrian detection [1] and vessel detection [2], significantly enhancing the safety in autonomous driving and marine traffic. In this work, we applied this idea to identifying atomically-thin semiconductor materials, and explore approaches to improve the performance by integrating visible and photoluminescence images.

Transition-metal dichalcogenide (TMDC) is a family of layer-structured materials, such as $MoS_2$ and $WSe_2$. They can be prepared in the form of a single layer of atoms (monolayer) or devised few-layers [8], being the most commonly investigated two-dimensional (2D) semiconductor materials. In these forms, TMDC presents great optical and electronic properties, showing significant application promise as a novel material for semiconductor and optoelectronics. The preparation of monolayer is an essential step for the research and development with TMDC, but so far it commonly relies on post-selection by human. After mechanical exfoliation [9] from bulk crystals, experts can identify monolayer regions from other regions with multilayers and contamination with an optical microscope.

A useful property of monolayer TMDC is photoluminescence (PL), i.e. with proper optical excitation they can emit strong fluorescence at a certain wavelength that is robust to the stimulating wavelength. Few-layer and bulk structures have significantly weaker photoluminescence, thus with proper filtering and amplification, monolayers can stand out. Contamination (dubbed dirt in the following) left by the preceding procedures, however, can fluoresce at a similar level as monolayer. White-light illuminated images by a common RGB camera allow experts to distinguish dirt from the interested material. Correlating PL image with RGB image is a natural and common practice to identify monolayer regions, but it so far still requires heavy human involvement, therefore computer vision based approaches to processing PL and RGB images altogether are worthwhile exploring.

## 2. Related Work

Previous works have demonstrated the advantages and indispensability of multimodality object detection. The authors of reference [5] collected RGB and thermal IR image pairs of pedestrians under low light condition, and showed with their dataset LLVIP that the thermal IR images are more effective at pedestrian detection than the RGB images under these dark conditions. The state-of-the-art method on LLVIP is using the head of YOLOv5 with Cross-Modality Fusion Backbone [10]. As illustrated in Fig. 2 of Ref [10], the authors set up a two-stream architecture, each of which conducts feed-forward feature extraction on a single modality, and on top of that, they placed three transformers to blend the features into the other modality at multiple stages. The detection model with cross-modality fusion backbone has successfully achieved better performance than both single modality model, and the two-stream model without transformers.

Two dimensional material layer detection has been approached with deep learning, but as far as we searched, the previous works are based only on RGB images. In reference [4], the authors constructed a deep convolutional neural-network 2DMOINet to detect regions with different numbers of layer in various 2D materials. Although this model performs well in terms of identifying the bulk crystals and the background, the confusion between monolayer and 2-6 layer crystal is generally high. This behavior is reasonable with RGB-only monolayer detection due to the subtle difference in the optical contrast between background, monolayer and thin multi-layers, as exemplified in Fig. 1 .

Given the effectiveness of complementing common RGB images with thermal-IR on pedestrian detection, we believe that a multimodal approach with PL and white-light imagings can be powerful.

## 3. Dataset and Features

We troubleshot our dataloading pipeline and baseline models with the RGB-Thermal dataset LLVIP [5] for muiltimodal pedestrian detection. LLVIP is a recently released public dataset with paired visible and infrared images under dark traffic scenes. It includes 15488 pairs of images with annotated pedestrians. Each pair is aligned both spatially and temporally, thus it is ideal for training and image fusion. An example is shown in figure 8.

Since datasets including PL images of 2D materials are lacking, we created our own dataset, PLVIS2D, for training our muiltimodal monolayer detection model. Specifically, we sampled materials from multiple experiments, and for each region of the exfoliated material, we took both a white-light illuminated optical image and a PL image, using an optical microscope in Heinz Lab at Stanford.

In detail, we searched for a batch of $WSe_2$ flakes on Silicon-Silicon Dioxide wafers. For each exfoliation wafer, all fluorescing objects are recorded, and their corresponding optical images are taken. Example image pairs of a monolayer and a non-monolayer luminescent object are shown in 1. The imaging setup consists of a X-Y scanning stage, collection cameras and two light sources for the two imaging modalities.

The features in PL and optical RGB images are complementary. PL images have strong and distinctive features for potential monolayers while they are potentially from dirt and edges. On the other hand, in optical RGB images, monolayers have awful optical contrast because of they are only several-nanometer thick, but other fluorescent sources-like dirts or reflecting edges are more distinctive from the 2D material than in PL images. Thus we can refer to the optical images to cross-check the monolayer, distinguishing them from irrelevant fluorescence sources.

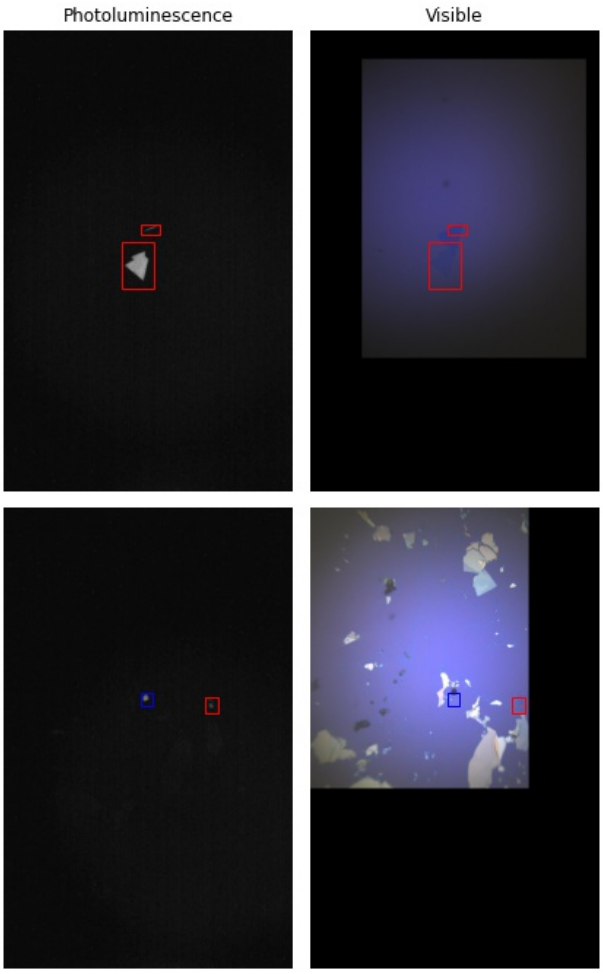


Figure 1. Example image pairs from PLVIS2D. The red boxes mark the monolayers, while the blue box marks the dirt.

This dataset includes 116 pairs of RGB-fluorescence images in total, which are split in to train/validate/test datasets of 92/12/12 pairs respectively. Since the spatial scale and center position are not aligned between the two imaging modalities, the coordinates of the scanning stage are also recorded and used to align the RGB and fluorescence images. With a global function that linearly maps the stage coordinates to a shift+scaling transformation, we aligned the RGB image to the PL image in each pair. Due to the stage's imperfection such as hysteresis, the two images are only weakly aligned. Temporal alignment is not necessary because the objects in PLVIS2D stay still. We then manually labeled all the monolayers and dirt.

Another notable feature of PLVIS2D as a dataset, is the flexibility for data augmentation. The objects are physically invariant under rotation and flipping, and scaling can be regarded as zooming in or out with different objectives on the microscope. The dataset should also be insensitive to small shearing, because the 2D material flakes are amorphous on the length scale of the images, and the imaging setup could be slightly tilted from the perpendicular direction to the plane. Therefore, we pre-augmented the dataset with random rotation within $\pm 90°$, random scaling within $\pm 10\%$ and shearing, and random shearing within $\pm 10°$. In addition, we randomly sampled $\sim 80\%$ of all the original images to cut out the object and obtain background images. All transformations are conducted on pairs of PL and aligned visible images, and the augmentation for the train/val/test sets are completely separated. After pre-augmentation, PLVIS2D is at the size of 996/103/79 for the train/val/test data sets, respectively, among which there are $\sim 5\%$ background images.

# 4. Methods

## 4.1. RGB + thermal IR Pedestrian detection

### 4.1.1 Baseline models

Firstly, we repeat the baseline models described in [5]. Similar to the paper, we base on YOLOv5, a single-stage object detection network, to develop our models. Although single-stage models are generally less accurate than two-stage networks, they are more light weight and much faster. Since LLVIP is a large dataset, we choose the large network in YOLOv5.

The baseline models here are basically the pretrained YOLOv5 networks fine tuned on each modality respectively. There are 499 layers, with 47 million parameters. We obtained the weights of these two baseline models from [5] and validated their performance.

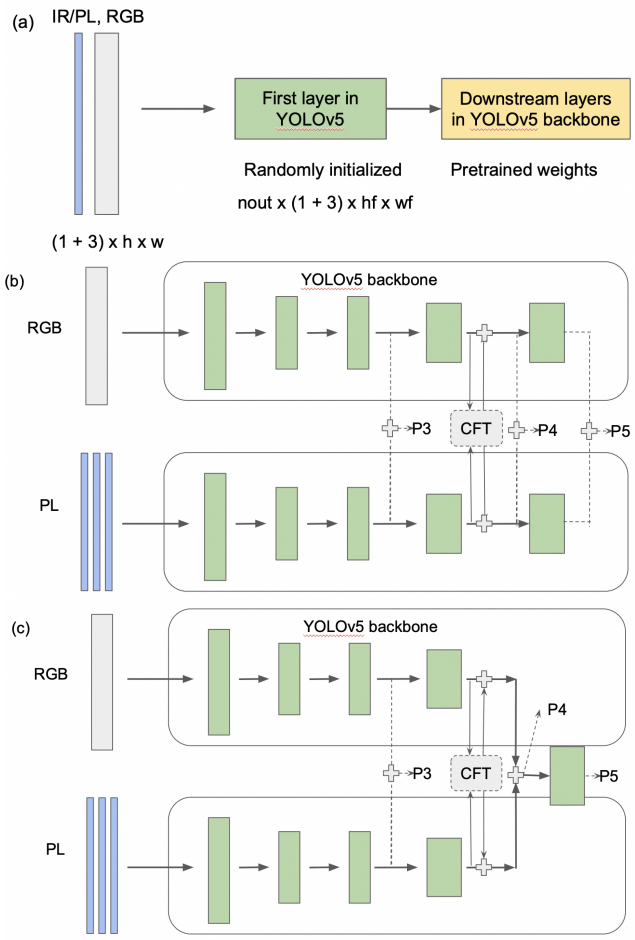


Figure 2. Illustration of the backbone architectures in multimodal models. All the models share the same detection head as YOLOv5, and the head takes feature maps at multiple stages, P3∼5, as input. Each annotation-free block represents a corresponding block of feed-forward layers in the YOLOv5 backbone [6], most of which are a convolution layer followed by a CSP bottleneck with 3 convolutions. (a) 4-channel model; (b) CFTx010: an experimental architecture modified from the proposed in Ref. [10], leaving only the middle transformer; (c) CFTx01fuse: another experimental Green blocks represents randomly initialization, and the yellow block represent pretrained weights.

### 4.1.2 Four-channel model

The most straightforward way to combine the information from different modalities is to concatenate them as complementary input channels at the first layer, which is the approach our four channel model takes.

As shown in figure 2(a), we convert the IR images to greyscale and concatenate it with the RGB images to form inputs of shape $4 \times h \times w$. The first layer of YOLOv5 is a focus layer, a variant of convolutional layer.

To accommodate the 4-channel inputs consisting of RGB

and thermal IR information, we load the model pre-trained on RGB images [5] and adapt the first layer from 3 input channels to 4. With the weights of this layer randomly initiated, we fine tune the model. After the modification on the first layer, this model has 2.3 thousand more parameters than the baseline models.

### 4.2. RGB + PL monolayer 2D material detection

For the multi-modal monolayer detection, we also trained two baseline models with the two modalities individually based on YOLOv5 [6]. Note that the PLVIS2D dataset is relatively small. To avoid overfitting, the nano network which scales the depth and width of each layer by 0.33 and 0.25 respectively is used throughout the experiments with this dataset.

Then we applied the four channel model to this dataset and used a genetic algorithm to tune the hyperparameters. Although the four channel model works naturally for the strictly aligned LLVIP dataset, we need to map our stage coordinates to the shift between the two modalities before concatenating the RGB + PL images to the four channel inputs. The mapping function was obtained by fitting the X, Y coordinates to the displacement between the center of the labeled objects in each RGB and PL image pair.

#### 4.2.1 Models with cross-modality fusion transformers in two-stream backbone

As mentioned in [3], the stage that cross-modality fusion is applied to the feature extraction stream matters to the application efficacy. Our four channel model concatenates the two modalities at the very first layer, which is effectively fusing the two images into one feature map at the output of the first layer and subsequently performing object detection on the fused feature map.

The authors of cross-modality fusion backbone [10] proposed to keep the feature extraction in a two-stream architecture, with three transformers blending the feature maps right before generating the features P3, P4, P5 that the detection head relies on. The transformers are allegedly able to correlate features extracted from one modality to the other and to enhance relevant features after referring to the other modality, which does not require them to be perfectly overlapped in space. We applied the exact cross-modality fusion transformer architecture [10], which will be referred as CFTx3 hereafter, to PLVIS2D. Then we did experiments on two variant architectures. The first one is with only the middle transformer out of the three, CFTx010, as illustrated in Fig. 2(b) . The second one, CFTx01fuse, takes the fused result at P4 for the downstream layers in YOLOv5's backbone.

As for the internal architecture of each fusion transformer, we use the same as Ref [10]. It contains a position encoding layer to integrate the features from the RGB and PL stream. Then a multihead self-attention layer is adopted to learn the correlation between the feature maps of two modalities, along with intra-modality correlation. The two outputs of a fusion transformer are added back to the two streams of extracted feature maps. The number of parameters in each transformer scales with $h^2 \times w^2$. Thus, to further trim the number of parameters in our model, we modified number of transformer embedded from three to one.

### 4.3. Data augmentation

YOLOv5 has integrated data augmentation functionalities in its data loader, and we applied it to all our training experiments. It includes mosaic combination, perspective transformation, horizontal/vertical flipping, and hue-saturation-value (HSV) scaling, all conducted stochastically according to the corresponding probability and/or range hyperparameters.

We used the default augmentation setting for the experiments on LLVIP. For our 2D material dataset, PLVIS2D, we applied random rotation in a large angular range in our own pre-augmentation and disabled the built-in rotation augmentation of YOLOv5, because the optional rotation augmentation in YOLOv5 is on the 2x2 mosaic rather than each image. The rotation in our pre-augmentation helps break the false correlation between the flake orientations in different panels of a mosaic, and it reduces the possibility of a flake or a dirt being clipped out of the mosaic. We also disabled the random HSV scaling, to avoid washing out the characteristic color patterns in the neighborhood of monolayer flakes in visible images. PL images have a single channel, so HSV is meaningless to them.

### 4.4. Evaluation

We evaluate the performance of the models using the common metrics for object detection: mean Average Precision, mAP@.5 and mAP@[.5,.95], corresponding to Intersection over Union (IoU) threshold at 0.5 and the average over a series of threshold values in [0.5,0.95], respectively. Since PLVIS2D has two classes, monolayer and dirt, and we value the performance of monolayer detection than dirt, we also present the AP@.5 and AP@[.5,.95] for monolayer, alongside the mAP's.

Additionally, we compare the models qualitatively in terms of their precision-recall curve, on validation sets. Recall is defined as the ratio of true-positive predictions to all actually positive samples, *i.e*. TP/(TP+FN), and precision is defined as the ratio of true-positive predictions to all positive predictions, *i.e*. TP/(TP+FP). An ideal object detector has its precision-recall curve close to a step function from one to zero at recall=1, meaning that precision and recall can be high at the same confidence level. For our models on the PLVIS2D dataset, we also compare them by the confu-

sion matrix, which describes the classification performance.

# 5. Experiments, Results and Discussion

## 5.1. Baseline models

The baseline model for monolayer detection is based on both the RGB and PL images respectively in PLVIS2D, with 3 channels as the input for the first layer), and then we evaluate their performance on the validation set.

We began with training the RGB-only model in the same architecture as YOLOv5n, with 3-channel jpg images as input. With fine-tuned hyperparameters and optimized weights from roughly trained results, the training results for 100 epochs is listed in table 1. Compared with PL-only and 4-channel models, RGB-only has relatively weak performance in mAPs, with around only 10.8% for mAP@.5 of monolayer.

As shown in the confusion matrix and PR curve in Fig. 3 and Fig. 6, the possibility for correctly predicted monolayers is only 12%. And the precision and recall cannot remain high at the same time, indicating that the RGB-only model has a relatively weak performance in identifying monolayers.

Then we trained the PL-only model with the same architecture. As shown in Table 1, it has better performance in mAPs compared with RGB only results, with around 66.2% mAPs for monolayers and 82.3% for all classes, partially because the contrast for PL images is much better than optical contrast.

The learning curve for PL-only model over 300 epochs is pretty flat as shown in Fig. 4. As indicated in the confusion matrix and PR curve in Fig. 3 and Fig. 6, the PL-only model can predict monolayer better than RBG-only model with probability of 75%. Moreover, the PR curve is closer to a step function, meaning that it can balance the precision and recall better.

## 5.2. Experiments

### 5.2.1 RGB+PL 4-channel

We use the default hyperparameters and weights provided by YOLOv5 to train the 4-channel model. The process over 300 epochs are shown in table 1 as 4-channel naive. From the mAP's of this model, we can see that it is already performing better than the baseline models. The mAP@0.5 for all classes are already close to 93%.

The confusion matrix and PR curve are also shown in Fig. 3 and Fig. 6, demonstrating excellent performance in all of classification, recall and precision.

Later, we attempt to finetune the hyperparameters with the built in genetic algoristhm in YOLOv5. Our baseline model for this specific hyperparameter tuning task is the 10 epochs training result of the model just discussed. The finetuned hyperparameters give smoother training curves than the previous model, shown in figure 5. However, it did not give significantly better results (see table 1 4-channel finetuned).

The RGB-only baseline model has a poor performance on detecting monolayers and quickly overfit, indicating that the generalization power of the model is insufficient, or lack of data. This is reasonable due to the subtle optical contrast a single atom layer can produce. The PL-only baseline model already shows much better performance due to the selectivity of PL imaging on monolayers. In addition, it also shows surprisingly good results at recognizing the monolayers from other dirt, which is not a trivial task for human eyes.

The 4-channel model further outperforms the PL-only baseline, suggesting the strength of combining multiple modalities. From the training results in figure 5 and the prediction on testing set, we can see that the model did not overfit within 300 epochs.

### 5.2.2 RGB+PL two-stream with fusion transformer

In light of the relatively satisfactory performance of the 4-channel model, we stick to the scaling factors of YOLOv5n in the experiments with CFT, which also makes the results fairer to compare with the 4-channel model's results.

We adapted the code in Ref. [10] to the structure of PLVIS2D, and trained a model with the original three-transformer architecture, CFTx3, using the same set of hyperparameters as the 4-channel naive model. Based on the validation performance, we terminated the training after 100 epochs, and the best result is listed in Table. 1 . Although it slightly exceeds the 4-channel finetuned model in validation mAP, its performance on AP@.5 and AP@[.5,.95] of monolayer has stagnated at values lower than the 4-channel naive model.

It is noteworthy that the 4-channel model has achieved reasonably good performance with around 3 million parameters, while CFTx3 has around 11 million parameters. We speculate that the complexity of CFTx3 is so high compared to the size of PLVIS2D that it needs meticulous hyperparameter tuning to reduce overfitting. Therefore, we set out to experiment with models that contain less complexity.

Removing the first and last transformer in CFTx3, we obtain CFTx010 as shown in Fig. 2(b). Trained with the same set of hyperparameters as the 4-channel naive model, it start to show a weak sign of overfitting after 200 epochs, thus we stopped training after 250 epochs. Its best validation results being listed in Table. 1, CFTx010 has generally outperformed the 4-channel model without finetuning the hyperparameters. The only exception is mAP@[.5,.95], which probably results in a compromised detection precision on dirt, because in terms of AP@[.5,.95] on monolayer it has exceeded the 4-channel model. Distinguishing

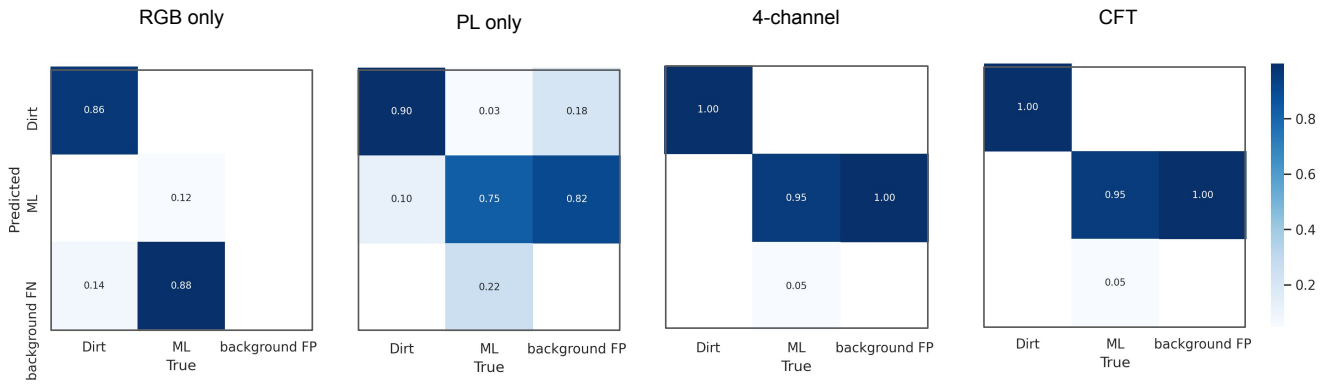


Figure 3. Confusion matrix of all the models.

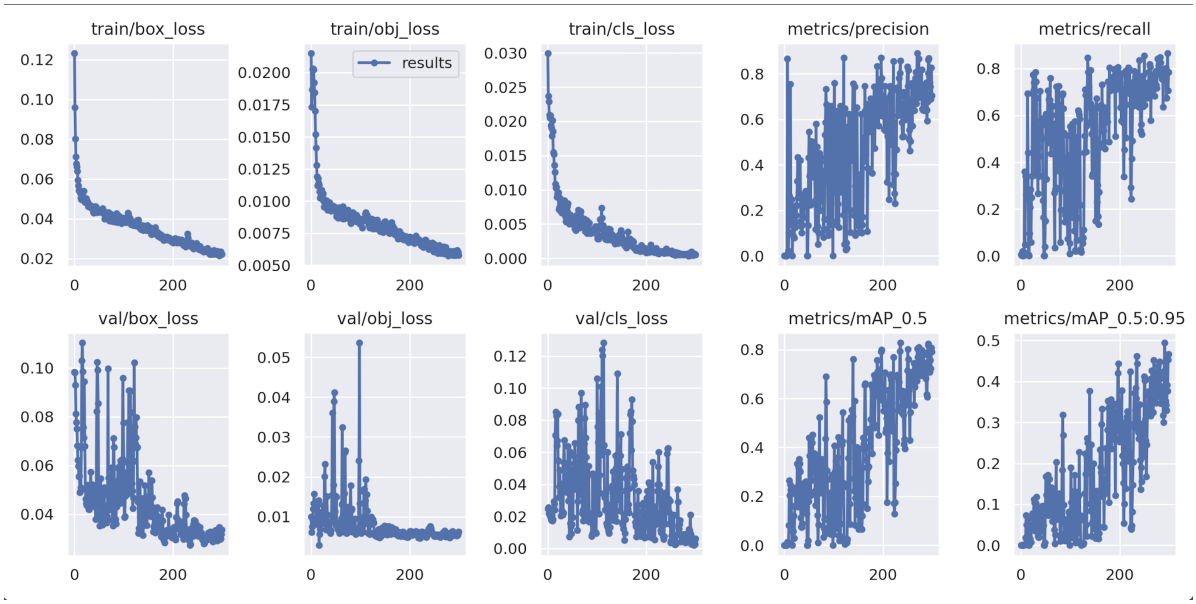


Figure 4. Learning process of the PL-only model after hyperparameter tuning.

dirt from monolayer is important, but high overlap with the ground-truth bounding box of dirt is unnecessary.

Inspired by the improvement from CFTx3 to CFTx010, we intuitively speculated that the feature maps at the stage of P4 are ready to be fused into a single stream. To test this hypothesis, we further modified CFTx010 into CFTx01fuse, by terminating the two-stream feed-forward starting at P4, and take P4 as the input to the rest layers in the backbone. The layers from P4 to P5 are unchanged, only reducing the number of stream from two to one. The best validation results of CFTx01fuse turn out to be the worst among the CFT models. Thus our hypothesis on merging architecture after P4 does not hold.

The best CFT model, CFTx010, is also validated on the test set, as shown in Table. 2 . It shows comparable performance with, if not better than, the other models, on the test set. Comparing the PR-curves shown in Fig. 6, we can also confirm that CFTx010 outperforms all the other models, because its PR-curve, especially the one associated with monolayer, has edged closer to the (1,1) corner.

## 5.3. Discussion

### 5.3.1 Complementariness of modalities

When testing our 4-channel model on LLVIP, we obtained better performance than the RGB-only model but worse than the IR-only model. On PLVIS2D, however, the 4-channel model easily out-performed both single-modality models. Without quantified investigation of the reasons, we argue that the additional complexity a multimodal model needs depends on the extent to which the images are com-

Table 1. Validation results comparison between models on PLVIS2D.

| Model | AP@.5 ML | AP@[.5,.95] ML | mAP@.5 | mAP@[.5,.95] |
|---|---|---|---|---|
| RGB-only | 0.108 | 0.0327 | 0.49 | 0.117 |
| PL-only | 0.662 | 0.330 | 0.823 | 0.495 |
| 4-channel naive | 0.894 | 0.389 | 0.945 | 0.541 |
| 4-channel finetuned | 0.857 | 0.404 | 0.926 | 0.536 |
| CFTx3 | 0.833 | 0.364 | 0.914 | 0.537 |
| CFTx010 | 0.927 | 0.424 | 0.961 | 0.519 |
| CFTx01fuse | 0.809 | 0.356 | 0.902 | 0.482 |

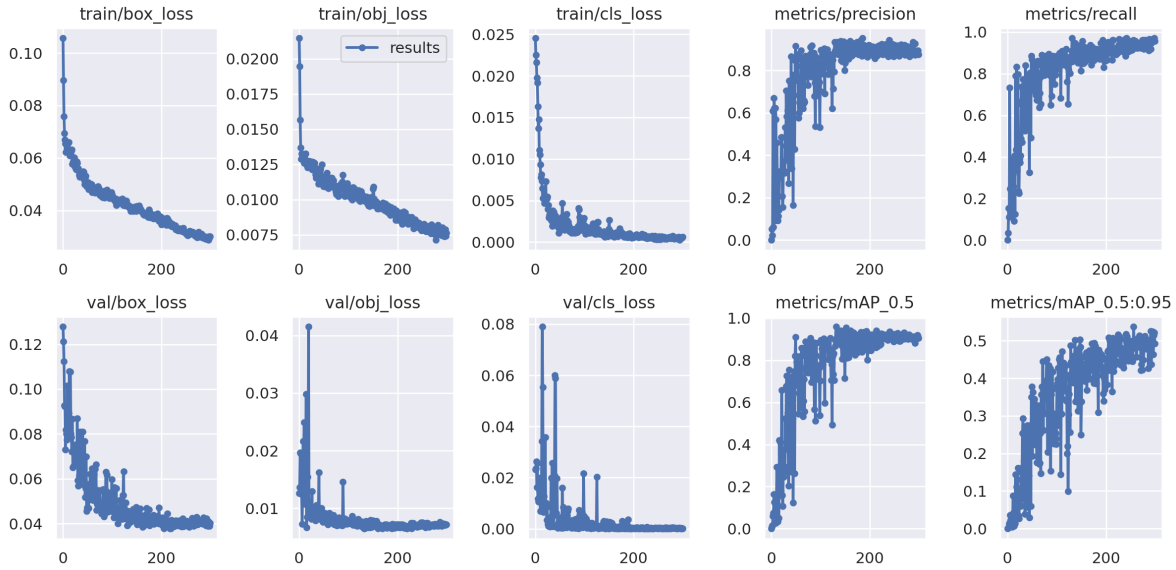


Figure 5. Learning process of the 4-channel model after hyperparameter tuning.

| Model | mAP@.5 | mAP@[.5,.95] |
|---|---|---|
| RGB-only | 0.304 | 0.1 |
| PL-only | 0.867 | 0.467 |
| 4-channel naive | 0.948 | 0.407 |
| 4-channel finetuned | 0.933 | 0.437 |
| CFTx010 | 0.969 | 0.508 |

Table 2. Test results comparison between models on PLVIS2D.

plementary to each other. When the features from different modalties are processed together, they can enhance or destructively blur features that are salient to a certain class of object, and model complexity determines the ability to coordinate them. The thermal IR images in LLVIP are indeed more informative than the corresponding RGB images at certain localities, as shown in Fig. 1 , and as shown by the better performance of the IR-only model than of the RGB-only model in Ref. [5]. When we introduce RGB images, inconsistency between visible and infrared images and surrounding objects could contribute to confusing features for the multi-modal detection model, and our 4-channel model turns out does not have sufficient complexity to handle them.

#### 5.3.2 Inaccuracy in manual labels in PLVIS2D

In the finetuned 4-channel model training process, the mAP stays relatively low while mAP0.5 entering a plateau. This possibly stems from the looseness of hand labeling. The labeled boxes are usually larger than the strict bounding box to accommodate slight shift between the image pairs, leading to difficulty for high-IoU predictions. Figure 7 compares the prediction with the labeling for an image pair the validation set, and clearly showed bounding boxes selected by our model that are more compact than the boxes in the original image and the pre-augmented image.

The large portion of background false positive is also possibly caused by missing labels. For some image pairs, the monolayer appears in the PL image but not the visible

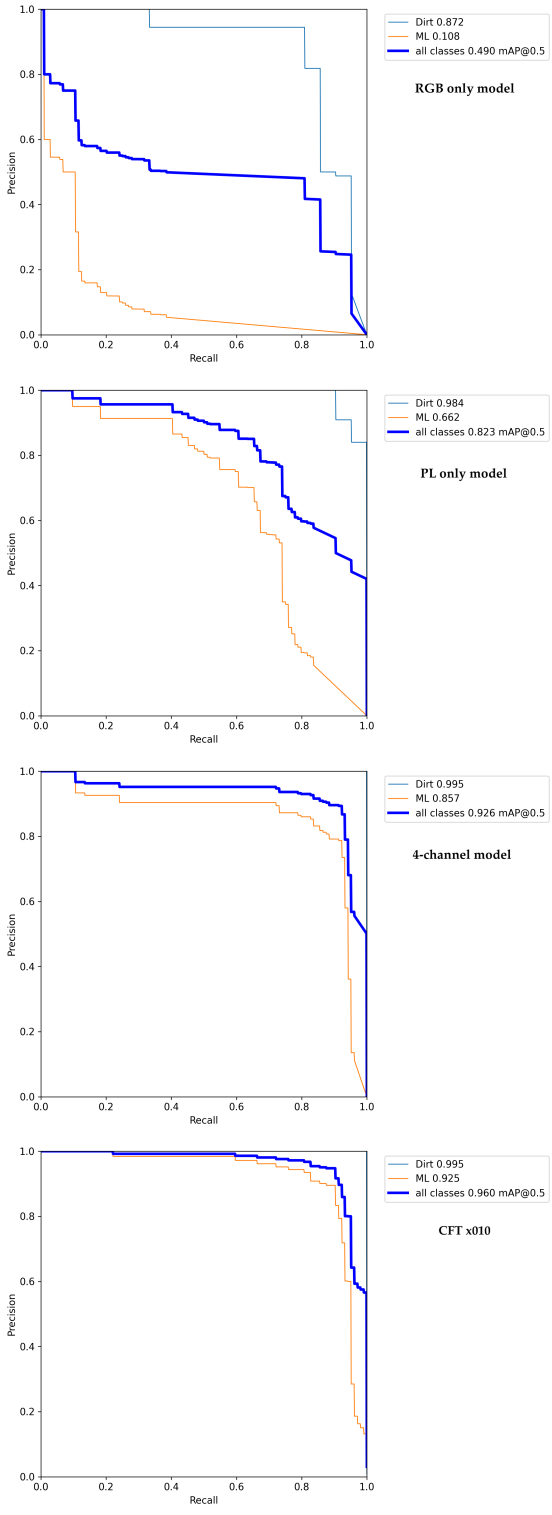


Figure 6. Precision-recall curves of the models.

image due to the difference in field-of-view of the two cameras. Another situation that are highly related to missing

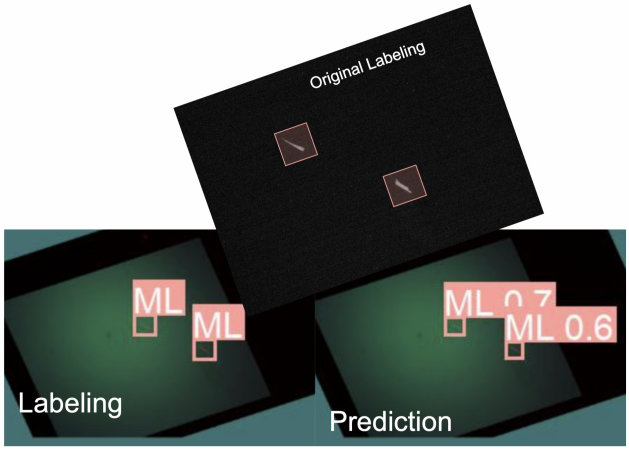


Figure 7. Labeling and prediction of of a image pair in the validation dataset, using the 4-channel model. For visualization purpose, the PL image is simply fused into the RGB image with 50% weight. The middle panel shows the labels in the original orientation of the image.

labels is the congestion of a group of monolayers in one vicinity. In these situations, it is difficult for human eye to discern a monolayer or to pick all the monolayers. Therefore, we only labeled the major flakes that appear in both images, which is practically enough for further operations. The fact that our model picks up all these human ignorance shows that it is already doing an amazing job.

With the progress of future experiments, our dataset, PLVIS2D, can be supplemented with increasing data, and our model can be trained and tested with the enlarged data set. In that case, manual labelling becomes an overhead that impedes the development of this procedure, so semi-supervised or self-supervised approaches can be valuable and promising future steps.

## 6. Conclusion

We experimented deep-learning based computer vision models that uses both PL and RGB images to identify where monolayer present in a region of exfoliated 2D semiconductor material. We collected a dataset, PLVIS2D, that contains weakly-aligned PL-RGB image pairs. Among the models we experimented on PLVIS2D, the 4-channel model outperforms both single-modal models, achieving satisfactory precision. The best model with a single cross-modality fusion transformer in the two-stream backbone and the same detection head as YOLOv5, further outperforms the 4-channel model. Our approach is a novel way to accelerate the automation of monolayer identification – a fundamental procedure in the field of 2D material research.

## 7. Supplementary Information

Visible image in LLVIP

Thermal IR image in LLVIP

Figure 8. Example image pairs from LLVIP.

## 8. Contribution and acknowledgements

J.W., J.H., and X.C. cooperated to complete the project for the data taking and processing, programming, analysis, and writing procedures. X.C. and J.H. prepared the dataset PLVIS2D. Specifically, they took the raw data together, and Jenny conducted the spatial alignment of images with stage coordinates and labelled the dataset. J.W. led the effort to adapt the existing code to interface the custom dataset with the routines for YOLOv5. Xueqi trained the single-modal baseline models. X.C. and J.W. implemented the 4-channel model, which J.H. trained and analyzed. J.W. adapted the fusion transformer code and trained the three CFT models. The experimental ideas on the variant CFT models (CFTx010, CFTx01fuse) is the product of team discussion. J.W. navigated the direction of this project. J.H. pioneers the analysis and led the writing. All members participated in the writing of this report.

We thank Dr. Feifei Li, Dr. Jiajun Wu, Dr. Ruohan Gao, and our TA Bohan Wu for their supportive suggestions and helps.

We would like to thank Mr. Twister for fruitful discussion, endless inspiration, and loving accompany throughout the whole journey.

## References

[1] M. Braun, S. Krebs, F. B. Flohr, and D. M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[2] F. Farahnakian and J. Heikkonen. Deep learning based multi-modal fusion architectures for maritime vessel detection. *Remote Sensing*, 12(16), 2020.

[3] D. Feng, C. Haase-Schutz, L. Rosenbaum, H. Hertlein, C. Glaser, F. Timm, W. Wiesbeck, and K. Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, mar 2021.

[4] B. Han, Y. Lin, Y. Yang, N. Mao, W. Li, H. Wang, K. Yasuda, X. Wang, V. Fatemi, L. Zhou, J. I.-J. Wang, Q. Ma, Y. Cao, D. Rodan-Legrain, Y.-Q. Bie, E. Navarro-Moratalla, D. Klein, D. MacNeill, S. Wu, H. Kitadai, X. Ling, P. Jarillo-Herrero, J. Kong, J. Yin, and T. Palacios. Deep-learning-enabled fast optical identification and characterization of 2d materials. *Advanced Materials*, 32(29):2000953, 2020.

[5] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou. Llvip: A visible-infrared paired dataset for low-light vision, 2021.

[6] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, J. Fang, imyhxy, K. Michael, Lorna, A. V, D. Montes, J. Nadar, Laughing, tkianai, yxNONG, P. Skalski, Z. Wang, A. Hogan, C. Fati, L. Mammana, AlexWang1900, D. Patel, D. Yiwei, F. You, J. Hajek, L. Diaconu, and M. T. Minh. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, Feb. 2022.

[7] H. Li and X.-J. Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019.

[8] K. F. Mak, C. Lee, J. Hone, J. Shan, and T. F. Heinz. Atomically thin $\mathrm{mos}_2$: A new direct-gap semiconductor. *Phys. Rev. Lett.*, 105:136805, Sep 2010.

[9] L. W. T. Ng, G. Hu, R. C. T. Howe, X. Zhu, Z. Yang, C. G. Jones, and T. Hasan. *2D Material Production Methods*, pages 53–101. Springer International Publishing, Cham, 2019.

[10] F. Qingyun, H. Dapeng, and W. Zhaokui. Cross-modality fusion transformer for multispectral object detection, 2021.