A Unified Multi-modal Structure for Retrieving Tracked Vehicles through Natural Language Descriptions

2023 CVPR AI City Challenge Track 2
Tracked-Vehicle Retrieval by Natural Language Descriptions

Dong Xie [1], Linhu Liu [1], Shengjun Zhang [2], Jiang Tian [1]

[1] AI Lab, Lenovo Research, Beijing, China
{xiedong2, liulh7, tianjiang1}@lenovo.com
[2] United Imaging Healthcare Surgical Technology, Wuhan, China
zsjcameron@gmail.com

Lenovo

Content

- Introduction
- Methodology
- Experiment
- Conclusion

Introduction

- The AI City Challenge Track 2 incorporates the language modality, called Natural language-based vehicle track retrieval. This task aims to retrieve single-camera tracks of vehicles that are consistent with the natural language query.

```
"nl": [
  "A red sedan drives forward.",
  "A red midsize sedan keep straight.",
  "A red car drove through an intersection."
],
"nl_other_views": [
  "A red sedan keeping straight.",
  "A red sedan runs down the street followed by a green van.",
  "A maroon sedan runs down the road followed by a green vehicle."
]
```

Figure 1. An example from CityFlow-NL for 2023 CVPR AI City Challenge Track 2.

Methodology

- An innovative deep learning system called Multimodal Language Vehicle Retrieval (MLVR) is developed for text-vehicle retrieval.

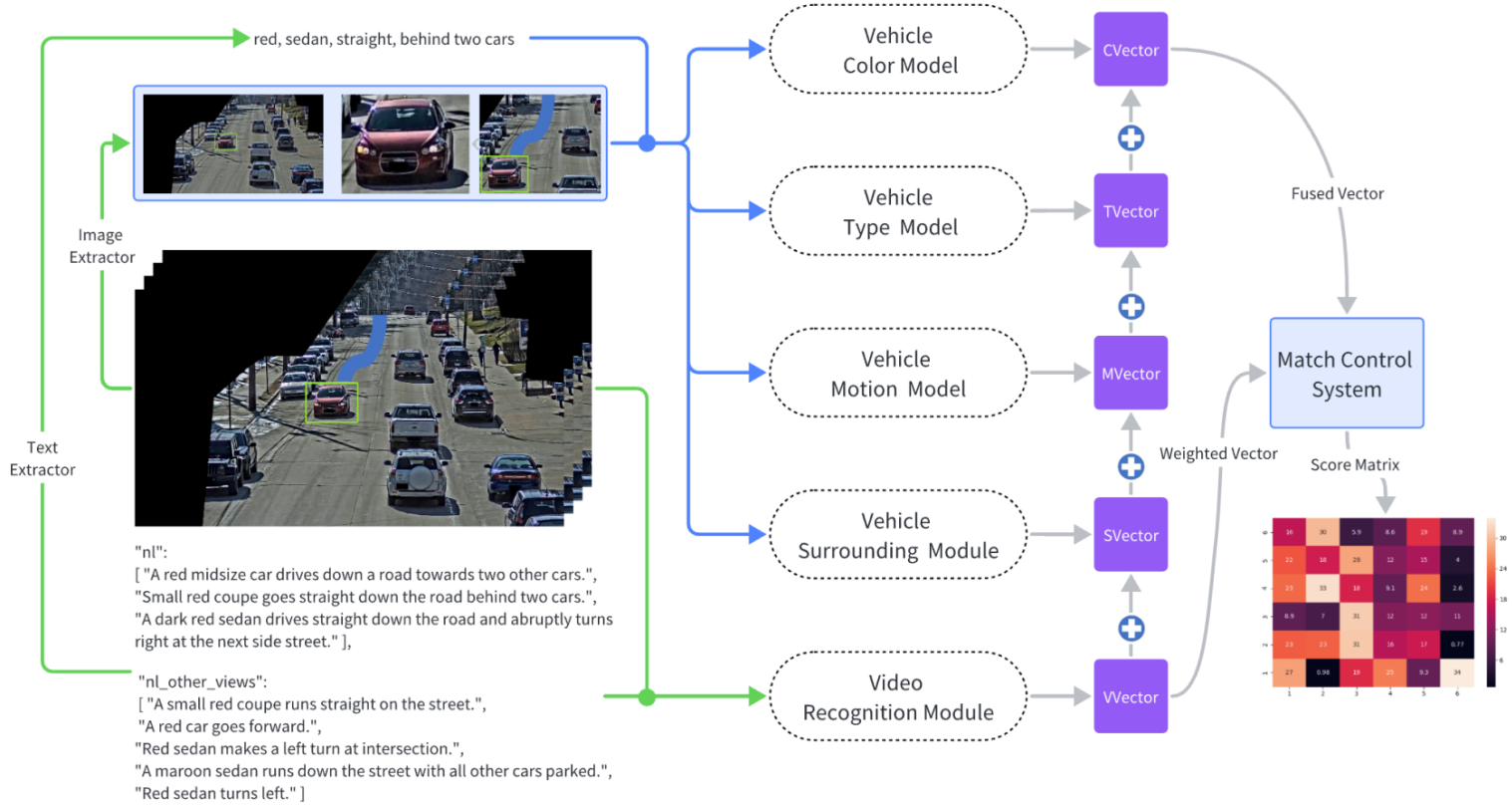

Figure 2. The structure of our MLVR system.

Methodology

- An analysis of Natural Language (NL) descriptions and corresponding descriptions from alternative perspectives (NL other view descriptions) reveals a connection.

Figure 3. The different video frames and NL descriptions of the same vehicle in the CityFlow-NL train dataset.

Methodology

- The video recognition module, which serves as the foundation of our MLVR model, is adapted from the X-CLIP algorithm to effectively discern the association between video clips and their corresponding text sentences.

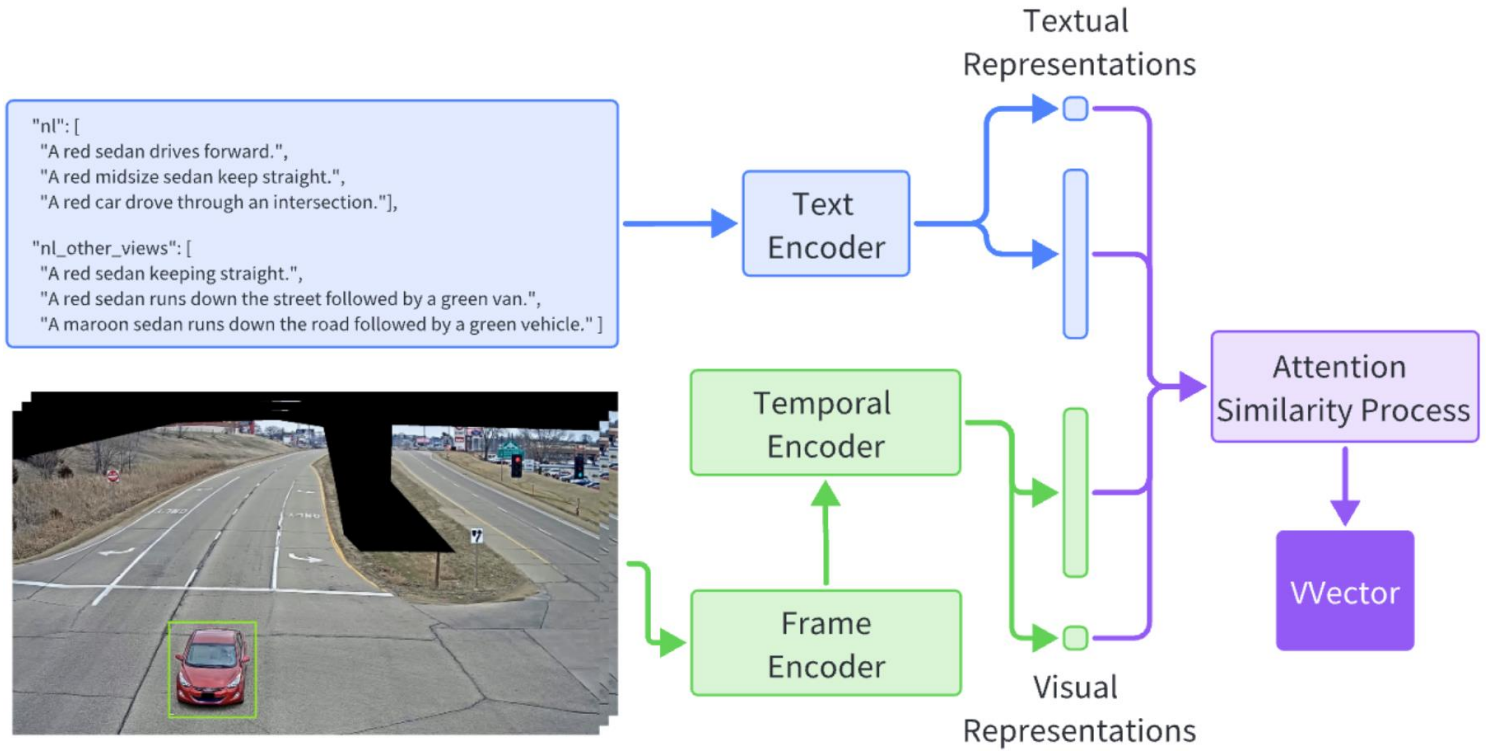

Figure 4. The primary architecture of the video recognition module.

Methodology

- The architecture of the vehicle color module, which employs a CLIP-based few-shot learning model, consists of several distinct segments.

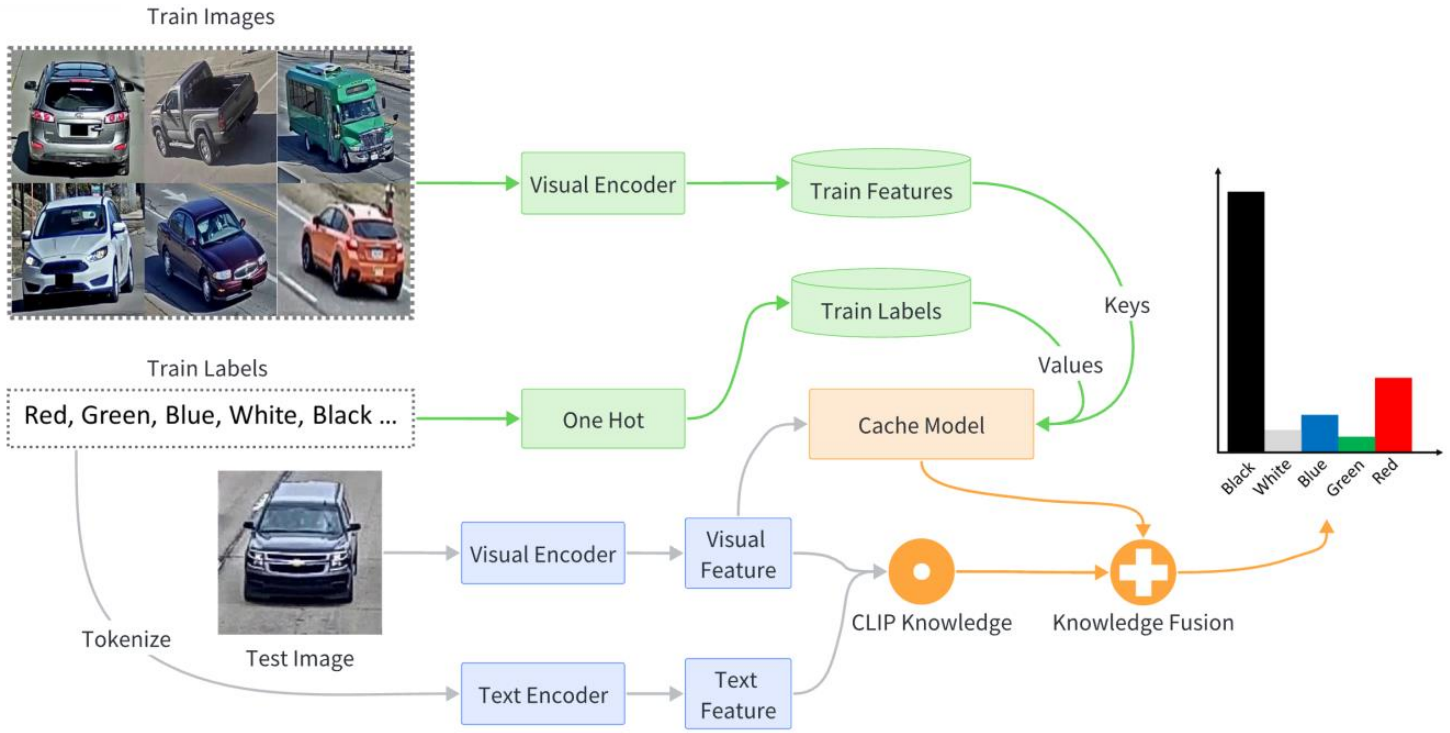

Figure 5. The architecture of the vehicle color module

Methodology

- Through an in-depth analysis of vehicle maneuver trajectories, the vehicle motion module has been developed as a cultured direction control system.

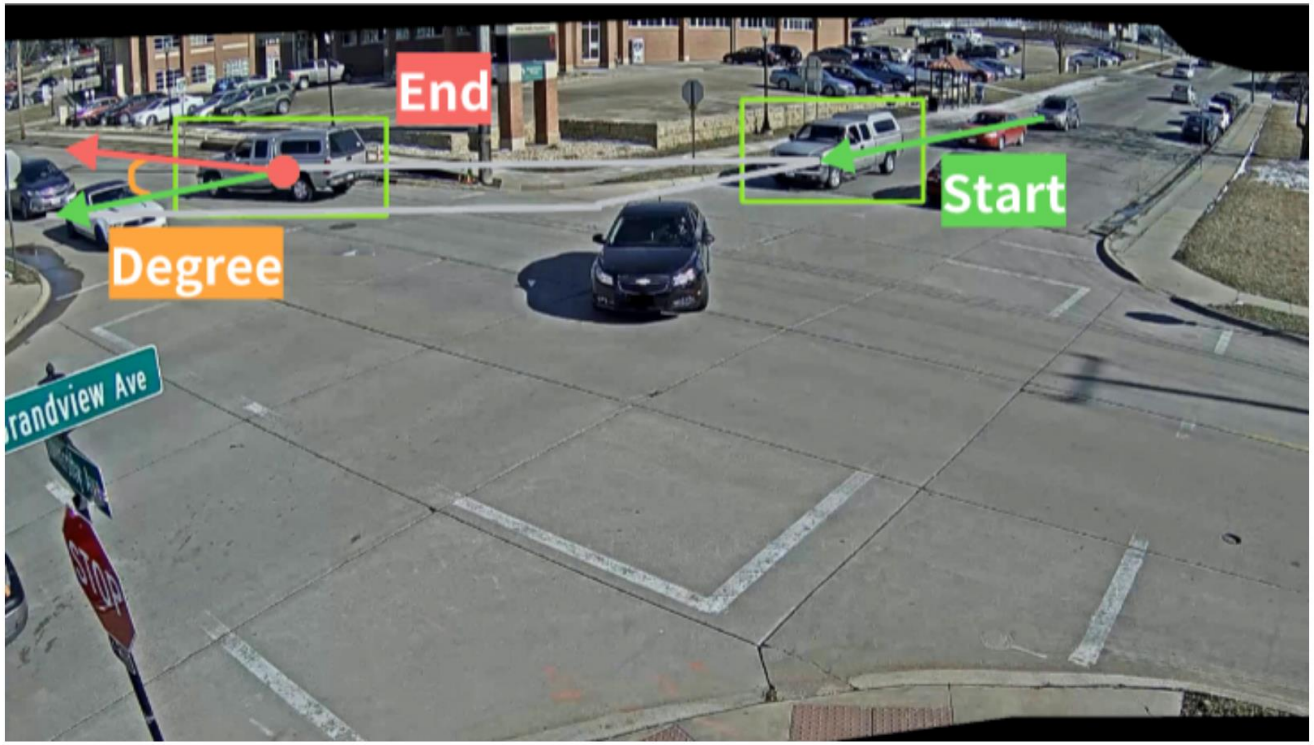

Figure 6. The example of vehicle motion module.

Methodology

- The vehicle surrounding module effectively leverages multiple sources of information to generate accurate predictions.

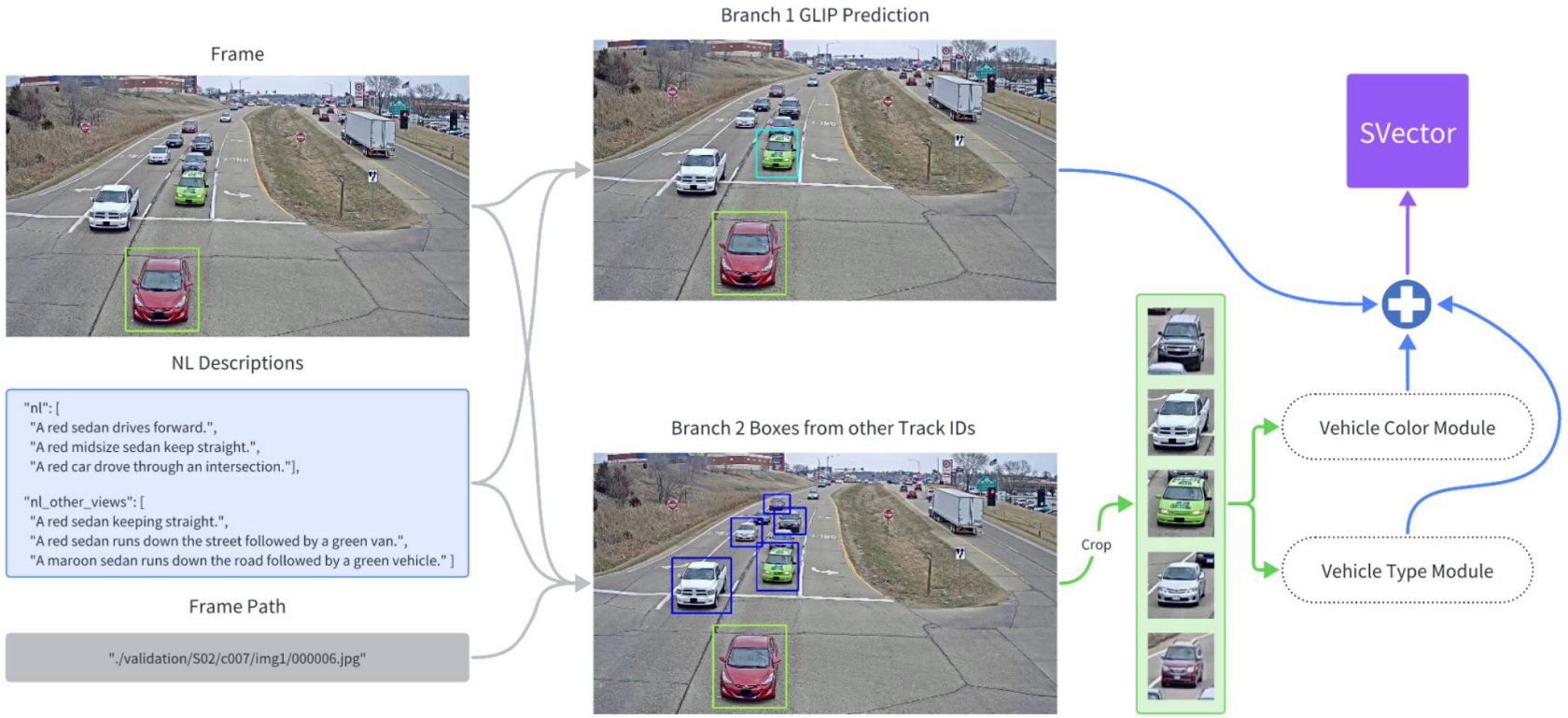

Figure 7. The structure of vehicle surrounding module.

Methodology

- Following vector data fusion, the match control system is employed to identify the optimal text-video match.

Algorithm 1 Matching Elimination System

1: Input the text-video matrix tv
2: **for** start row = 1, length **do**
3: Get the highest score column index hci in $tv[row, :]$
4: Get the highest score row index hri in $tv[:, hci]$
5: **if** $row == hri$ **then**
6: For every element in column hci except $tv[hri, hci]$, minus a threshold mt
7: **end if**
8: **end for**

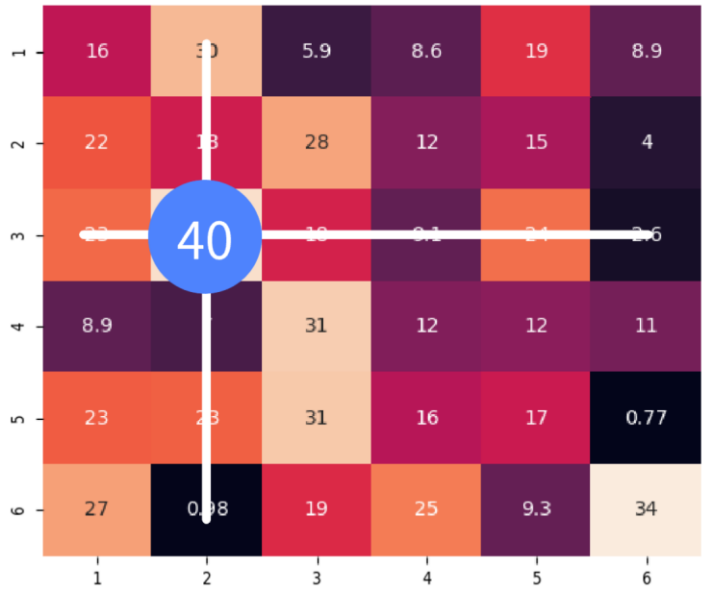

Figure 8. The matrix example of matching elimination system.

Experiment

- The dataset employed for the evaluation of the MLVR model is CityFlow-NL, consisting of 2,155 distinct vehicle trajectories and associated track IDs, as well as corresponding natural language descriptions.

- In addition to the primary dataset, a separate test set comprising 184 distinct vehicle trajectories is utilized to assess the MLVR model's final performance.

- The mean reciprocal rank (MRR) serves as the primary evaluation metric for assessing the performance of the MLVR model using the CityFlow-NL dataset.

Experiment

- The ablation study emphasizes the efficacy of each module in augmenting the overall performance of our MLVR model, and our MLVR model achieves a second-place ranking with an MRR score of 0.8179.

Baseline	VCT	VM	VS1	VS12	MCS	MRR
✓						0.2761
✓	✓					0.4191
✓	✓	✓				0.5885
✓	✓	✓			✓	0.6714
✓	✓	✓	✓		✓	0.7160
✓	✓	✓	✓	✓	✓	0.8179

Table 1. Ablation study analysis of our MLVR method.

Rank	Team ID	Team Name	MRR
1	9	HCMIU-CVIP	0.8263
2	**28**	**IOV**	**0.8179**
3	85	AIO-NLRetrieve	0.4795
4	151	AIO2022	0.4659
5	76	DUT_ReID	0.4392

Table 2. The public leaderboard of tracked-vehicle retrieval by natural language descriptions.

Conclusion

- Creation of the MLVR system, an innovative multimodal technique.
- MLVR uses text, image, and video data for enhanced vehicle tracking.
- Showcased exceptional performance in the 7th AI City Challenge.
- Demonstrated significant potential in traffic management.

References

- [1] 2023 AI City Challenge. https://www.aicitychallenge.org/
- [2] Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. Cityflow-nl: Tracking and retrieval of vehicles at city scale by natural language descriptions. arXiv preprint arXiv:2101.04741, 2021.
- [3] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In Proceedings of the 30th ACM International Conference on Multimedia, pages 638–647, 2022.
- [4] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10965–10975, 2022.

Thank You!

Q&A

Dong Xie

Email: xiedong2@lenovo.com

Code: https://github.com/eadst/MLVR

Lenovo