# MVRNet: A Multi-View Fusion Network for Road Extraction from Satellite and UAV Imagery


1st Wenfei Zhao
*Hunan University*
Changsha, China
zwf24@hnu.edu.cn

2nd Juan Luo*
*Hunan University*
Changsha, China
juanluo@hnu.edu.cn

3rd Kexuan Feng
*Hunan University*
Changsha, China
kexuanfeng@hnu.edu.cn

4th Ying Qiao
*Hunan University*
Changsha, China
qy2020@hnu.edu.cn

5th Shuyang Teng
*Hunan University*
Changsha, China
tsy0603@hnu.edu.cn



***Abstract*—Accurate road extraction from remote sensing imagery serves as a critical data foundation for urban and rural planning and intelligent transportation systems. However, existing methods based on single view often struggle with missed detections and fragmented road networks caused by occlusions and shadows. To address these challenges, this paper constructs a paired satellite-UAV road extraction dataset called SURoad and proposes a multi-view road extraction network named MVRNet. The proposed method adopts a layered collaborative modeling mechanism, which organically integrates shallow-level multi-view feature alignment for detail enhancement with deep-level satellite topological priors for semantic guidance. Through an adaptive weighting strategy, the model achieves adaptive integration modeling of global structural consistency and local road details. Experiments on the SURoad dataset show that MVRNet outperforms existing methods on key metrics such as F1 score and IoU, surpassing the best baseline by approximately 1.3% in IoU. These results confirm the effectiveness of the layered collaborative modeling strategy in handling complex occlusions and feature misalignments.**



***Index Terms*—road extraction, multi-view fusion, layered collaborative modeling, satellite and UAV imagery**


## I. Introduction

Accurate extraction of road networks is a crucial foundation for intelligent transportation applications such as urban planning, autonomous driving navigation, and emergency response logistics [1]. Although deep learning techniques [2], [3], [4]have shown strong capabilities in automated road extraction tasks, existing methods often rely on remote sensing observations from a single perspective. Dense tall buildings or heavily vegetated areas create blind spots in the physical view [5], resulting in broken roads or blurred road shapes in the extraction results.

In response to these perception bottlenecks, multi-source information fusion has been introduced to improve both the structural modeling and geometric consistency of complex road networks. For instance, work [6] utilizes the sensitivity of SAR signals to surface textures, effectively recovering missing road features in optical blind spots. Study [7], combined with the height information provided by LiDAR point clouds and the physical characteristic that roads are typically on the ground, helps the model distinguish between elevated building rooftops and lower road surfaces. Yuan et al. [8] integrated crowdsourced GPS trajectory data as auxiliary constraints to introduce spatial distribution priors of roads into the model learning process. However, these methods rely on additional sensors or external data sources, often suffering from high costs or noise interference.

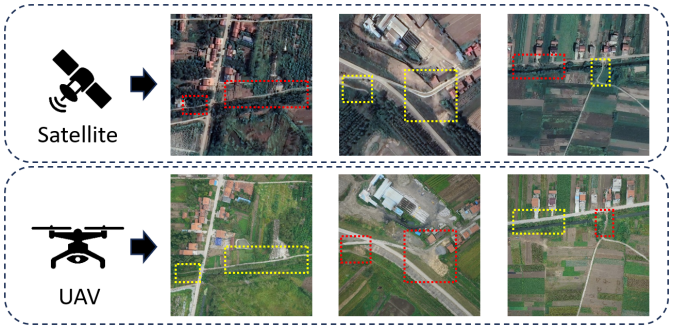


Fig. 1. Comparison of road visibility under different remote sensing views. The top row shows satellite images, and the bottom row shows UAV images. Yellow dashed boxes indicate clear road structures, while red dashed boxes denote blurred regions.

As Space-Air-Ground collaborative perception gradually becomes a mainstream research paradigm, cross-view collaboration between satellites and UAVs provides a more feasible approach to address these issues, taking advantage of its significant low cost and complementary benefits. As shown in Fig. 1, satellite imagery excels in maintaining global connectivity [9] but is limited in capturing detailed boundaries. In contrast, while UAV imagery clearly captures local road details [10], it suffers from structural discontinuities due to a lack of spatial context. The significant complementarity between the strengths and weaknesses of these two views provides a solid foundation for multi-view collaborative modeling. Existing studies further support the feasibility of cross-view modeling. For instance, Zheng et al. [11] proposed a Siamese network to map images from different viewpoints into a shared feature space to learn semantic representations invariant to viewpoint changes. Research by Regmi et al. [12] shows that during cross-view spatial transformations, the geometric structure of scenes such as roads is well preserved.

However, collaborative road extraction using satellites and UAV still faces significant challenges. As a pixel level segmentation task, it requires precise feature alignment. Yet, the significant differences in imaging geometry and resolution between images from different perspectives make it difficult to precisely align cross-view features. Meanwhile,

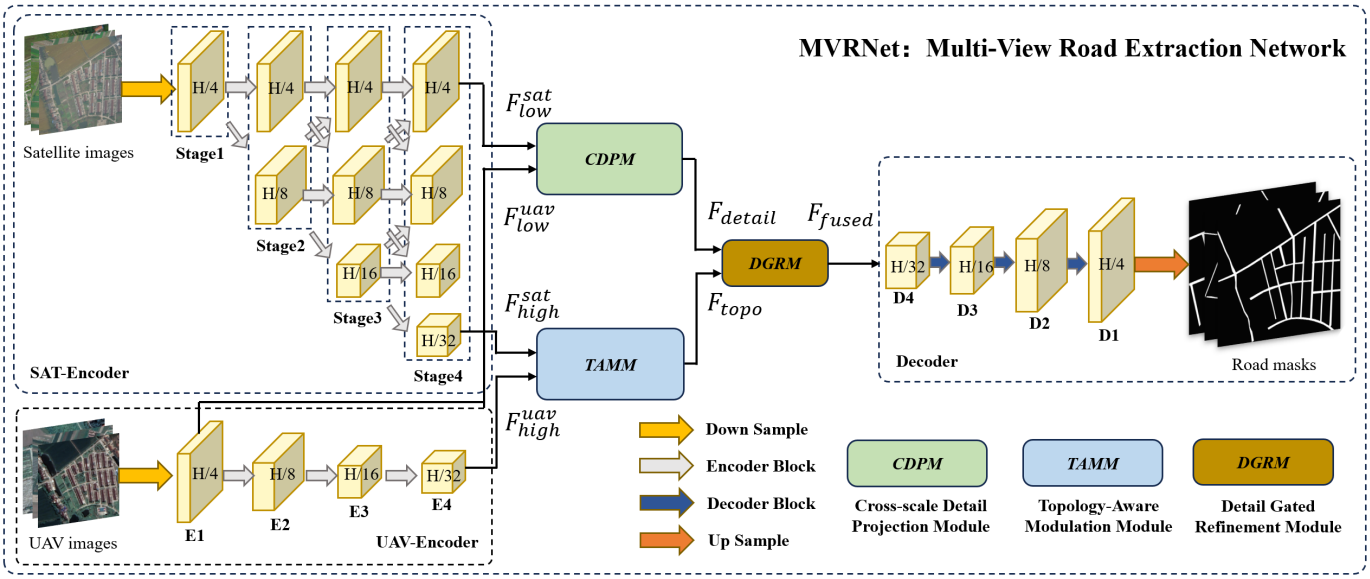


Fig. 2. The overall architecture of the proposed MVRNet.

under conditions where alignment errors exist, it remains a significant challenge to effectively model and integrate their complementary characteristics in both macro topology and local details.In addition, the lack of paired multi-view annotated datasets with high precision also limits progress in this field. To address these issues, this study proposes a Multi-View Road Extraction Network, termed **MVRNet**, which aims to fully explore and integrate the complementary advantages of both sources through collaborative modeling. The main contributions of this paper are as follows:

- We propose a layered collaborative modeling strategy to tackle cross-view fusion challenges. Specifically, we design a Cross-view Detail Projection Module (CDPM) to correct local misalignments in shallow layers via **sparse constrained attention**, and a Topology-Aware Modulation Module (TAMM) to inject satellite connectivity priors into deep layers via **adaptive affine transformations**, effectively resolving spatial and topological inconsistencies.
- We developed a Detail Gated Refinement Module (DGRM) to balance structural stability with precise details. By employing an **adaptive gating mechanism** at the pixel level, the module dynamically weights deep and shallow features, enabling precise restoration of road shapes while preserving global network connectivity.
- We constructed **SURoad**, a heterogeneous satellite-UAV cross-view dataset for road extraction, comprising 1,479 strictly registered image pairs. This dataset addresses the scarcity of reliable paired observational data in existing research and provides a solid data foundation for multi-view feature fusion and collaborative modeling.

## II. Proposed Method

This section details the implementation of MVRNet, which is designed as a multi-view fusion network for road extraction from satellite and UAV imagery. As illustrated in Fig. 2, the network uses dual-branch encoders to extract features at multiple scales and achieves collaborative modeling at different levels through three key modules. Specifically, the CDPM aligns different views in shallow layers to generate representations with enhanced details; the TAMM introduces satellite structural information in deep layers to form topological constraints; the DGRM adaptively fuses the features.

### A. The Dual-branch Encoder of MVRNet

We employ differentiated backbone networks tailored to the characteristics of different data sources. As shown in Fig. 2, the satellite branch uses lightweight HRNet [13] to accurately capture the global topology and geometry of the road by taking advantage of its multi-resolution parallel interaction characteristics. Meanwhile, the UAV branch employs ResNet-34 [14] to effectively represent textures in close range views.

For feature construction, we extract shallow and deep representations from Stage 4 of HRNet ($F_{\text{low}}^{\text{sat}}, F_{\text{high}}^{\text{sat}}$) and layers E1/E4 of ResNet-34 ($F_{\text{low}}^{\text{uav}}, F_{\text{high}}^{\text{uav}}$). To address the heterogeneous channel sizes, all extracted features are projected to unified dimensions via $1 \times 1$ convolutions prior to cross-view fusion.

### B. The CDPM of MVRNet

Fine road structures rely on high resolution features for representation. However, direct fusion often leads to edge blurring due to the spatial misalignment caused by viewpoint differences. To address this, the CDPM is proposed as illustrated in Fig. 3. By explicitly modeling geometric correlations, this module achieves precise alignment and injection of satellite context information while preserving the dominance of UAV details.

Specifically, given the channel-projected and aligned shallow features $F_{\text{low}}^{\text{sat}}$ and $F_{\text{low}}^{\text{uav}} \in \mathbb{R}^{C_{\text{low}} \times H/4 \times W/4}$, where $C_{\text{low}} =$

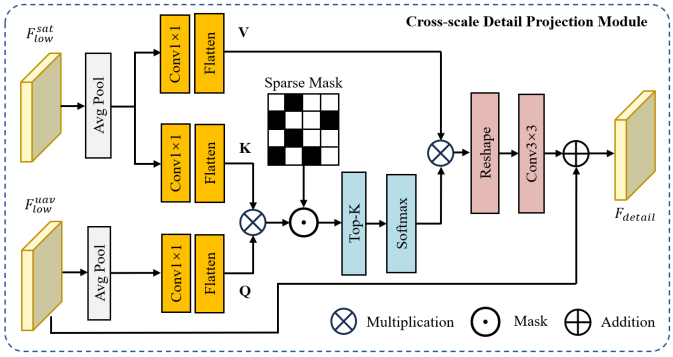


Fig. 3. Detailed structure of the CDPM.

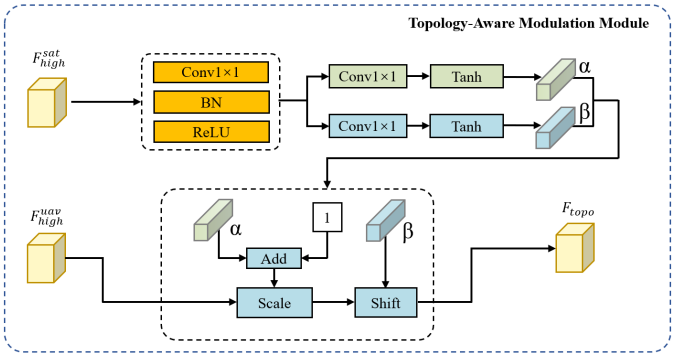


Fig. 4. Detailed structure of the TAMM.

64, and $H, W$ denote the height and width of the input images respectively. To reduce the computational redundancy caused by high-resolution shallow features, we first employ average pooling to spatially downsample the input features, followed by a convolution to map them into a subspace. To correct the deviation of satellite features based on the UAV view, we construct a cross-view attention mechanism using $F_{\text{low}}^{\text{uav}}$ as the Query, and $F_{\text{low}}^{\text{sat}}$ as the Key and Value.

Unlike standard global attention, which assumes full-map correlations, CDPM introduces a sparse mask and a Top-K selection strategy to eliminate invalid matches. Specifically, we first construct a mask based on the locality principle, enforcing that Query feature points $(i, j)$ only interact with feature points within a local neighborhood centered at $(i, j)$ with a radius of $R$ in the Key feature map, directly blocking long-range correlations outside the window. On this basis, we further perform Top-K selection on the correlation scores within the window, retaining only the top $K$ response values. This process encourages the network to focus on regions that are geometrically consistent and well aligned in texture. The above process and the final attention aggregation can be expressed as:

$$O = \text{Softmax}\left(\text{TopK}\left(\frac{QK^{\top}}{\sqrt{d}} \odot M_{\text{sparse}}\right)\right) V \tag{1}$$

where $M_{\text{sparse}}$ denotes the sparse mask. The TopK$(\cdot)$ operation retains only the top $K$ maximum values and sets the others to negative infinity.

The aggregated feature $O$ is restored to the original spatial dimensions through Reshape and convolution, obtaining the aligned feature. Finally, it is injected into the original UAV feature stream through a residual connection to output the final feature:

$$F_{\text{detail}} = F_{\text{low}}^{\text{uav}} + \text{Conv}_{3\times3}(\text{Reshape}(O)) \tag{2}$$

This design not only retains the texture details of the UAV images but also effectively integrates the geometrically corrected complementary satellite information.

### C. The TAMM of MVRNet

Although CDPM improves the precision of boundary details, constructing the integrity of the road network requires macroscopic topological guidance. To this end, the TAMM is designed at the deep semantic scale to constrain UAV topological features using the macro structural priors of satellite imagery.

Let the deep-scale features of the UAV and satellite be $F_{\text{high}}^{\text{uav}}$ and $F_{\text{high}}^{\text{sat}}$ respectively, where $F_{\text{high}}^{\text{uav}}, F_{\text{high}}^{\text{sat}} \in \mathbb{R}^{C_{\text{high}} \times H/32 \times W/32}$ and $C_{\text{high}} = 256$. As shown in Fig. 4 , to extract pure topological information from satellite features, a structure extraction unit is employed to capture directional patterns related to road connectivity while suppressing background noise:

$$T = \text{ReLU}\big(\text{BN}\big(\text{Conv}_{1\times1}\big(F_{\text{high}}^{\text{sat}}\big)\big)\big) \tag{3}$$

Based on the structural features $T$, the FiLM mechanism is introduced to generate channel level control parameters. The scaling factor $\alpha$ and bias $\beta$ are predicted via two separate convolutions followed by Tanh activation:

$$\begin{aligned} \alpha &= \text{Tanh}(\text{Conv}_{1\times1}(T)), \\ \beta &= \text{Tanh}(\text{Conv}_{1\times1}(T)). \end{aligned} \tag{4}$$

Here, Tanh constrains the parameters within $[-1, 1]$ to ensure gradient stability. Finally, the generated parameters are utilized to perform an affine transformation on $F_{\text{high}}^{\text{uav}}$:

$$F_{\text{topo}} = (1 + \alpha) \otimes F_{\text{high}}^{\text{uav}} + \beta \tag{5}$$

This design successfully incorporates the global topological prior of satellites while preserving the discriminative semantics of UAV.

### D. The DGRM of MVRNet

Based on the complementarity between the structural semantics of $F_{\text{topo}}$ and the high-frequency details of $F_{\text{detail}}$, this paper propose DGRM, as shown in Fig. 5. By introducing pixel-level gating to dynamically balance the two types of features, the model selectively restores detailed information under structural constraints.

First, $F_{\text{detail}}$ is upsampled to the deep feature scale and processed by a feature mapping unit to generate the candidate detail feature:

$$F_{\text{det}} = \text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(\text{Up}(F_{\text{detail}})))) \tag{6}$$

where $F_{\text{det}} \in \mathbb{R}^{C_{\text{high}} \times H/32 \times W/32}$.

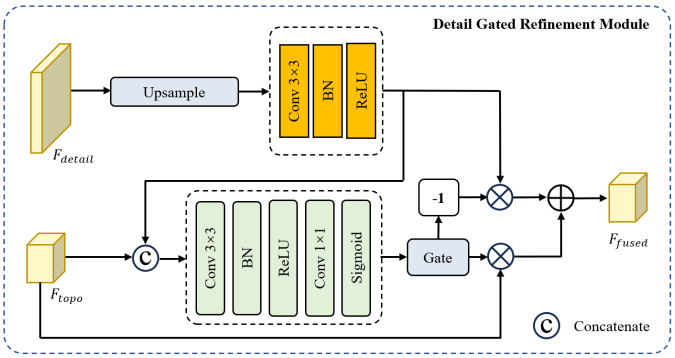


Fig. 5. Detailed structure of the DGRM.

Next, the topological feature $F_{\text{topo}}$ and the candidate detail feature $F_{\text{det}}$ are concatenated along the channel dimension and generate the pixel-level gating map $G$ using a convolutional network and a Sigmoid function:

$$G = \sigma(\text{Conv}_{3\times3}(\text{ReLU}(\text{Conv}_{1\times1}([F_{\text{topo}}, F_{\text{det}}])))) \quad (7)$$

where $[\cdot,\cdot]$ denotes channel-wise concatenation, $\sigma(\cdot)$ denotes the Sigmoid activation function..

Finally, a complementary gating strategy is applied to fuse the features:

$$F_{\text{fused}} = G \otimes F_{\text{topo}} + (1 - G) \otimes F_{\text{det}} \quad (8)$$

This mechanism prioritizes the stability of deep-level topology in the main areas, while automatically switching to guidance from shallow-level details in finer branches. This dynamic interplay successfully resolves the conflict between semantic integrity and boundary accuracy, achieving pixel-level precise restoration of complex road networks.

### E. Loss Functions Used in MVRNet

To address the class imbalance caused by the sparsity of roads in remote sensing images, this paper constructs a hybrid loss function weighted by BCE and the Dice Similarity Coefficient (DSC). BCE aims to constrain pixel-level classification accuracy, while Dice loss measures set similarity at the regional level to alleviate sample imbalance and enhance geometric connectivity. The loss is defined as follows:

$$L_{\text{bce}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1 - y_i)\log(1 - p_i)] \quad (9)$$

$$L_{\text{dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i p_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} p_i + \epsilon} \quad (10)$$

$$L = L_{\text{bce}} + L_{\text{dice}} \quad (11)$$

Here, $N$ denotes the total number of pixels while $y_i$ and $p_i$ represent the ground truth label and predicted probability respectively. The symbol $\epsilon$ is the smoothing term.

## III. SURoad Dataset

The SURoad dataset is constructed to facilitate road extraction research under satellite and UAV viewpoints. In the following, we describe the dataset construction process in detail.

### A. Dataset Construction

The raw data are derived from the **UAV-VisLoc** dataset [15], which includes satellite orthophotos with a ground sampling distance of 0.3 m, UAV image sequences with resolutions ranging from 0.1 m to 0.2 m, as well as corresponding metadata. To achieve strict spatial alignment we perform geographic registration based on metadata. First, the center point of the UAV image on the satellite base map was determined via coordinate projection. Second, construct a rotation matrix based on the coverage area and heading angle to crop the region aligned with the UAV's field of view. Finally the aligned pairs are center cropped and resampled to 512 by 512 pixels using bilinear interpolation.

### B. Image Annotation

Since satellite images have low resolution and serve only as reference, we generated labels only for UAV images. We adopted manual annotation using the Labelme tool. Experts carefully reviewed and corrected the results to obtain precise binary road labels.

### C. Dataset Statistics

The SURoad dataset consists of 1479 strictly aligned satellite–UAV image pairs, uniformly cropped to 512 × 512 pixels. It covers typical road scenes from urban to rural areas and includes challenging cases with occlusion and shadow, supporting the evaluation of road detail recovery and generalization.

## IV. Experiment

### A. Implementation Details

All experiments are conducted on the SURoad dataset, which is partitioned into training, validation, and test sets in an 8:1:1 ratio. The model is implemented using the PyTorch framework on a single NVIDIA GeForce RTX 4090 GPU. The maximum number of training epochs is set to 150, with a batch size of 4. We adopted an early stopping strategy with a patience of 15 epochs based on the F1 score on the validation set. The Adam optimizer was used with a learning rate of $3 \times 10^{-5}$. For the CDPM module, the parameters are set to $R = 4$ and $K = 20$.

### B. Comparative Experiments

To verify the effectiveness of the proposed MVRNet, we selected U-Net [16], DeepLab v3+ [17], D-LinkNet [18], NL-LinkNet [19], RoadFormer [20], StripUnet [21] and CGCNet [22] as comparison baselines. Given that the above baseline methods all use a single-encoder structure, the experiments uniformly use UAV images and their corresponding annotations as input to ensure the fairness of the comparative experiments. At the same time, this paper uses Precision (P), Recall (R), F1-Score (F1), and Intersection over Union (IoU) as quantitative evaluation metrics.

Table I quantitatively presents the performance of various models on the test set. MVRNet significantly outperforms the baseline models across all core metrics, achieving 83.27%, a Recall of 83.06%, and an F1-score of 83.17%. Notably, in

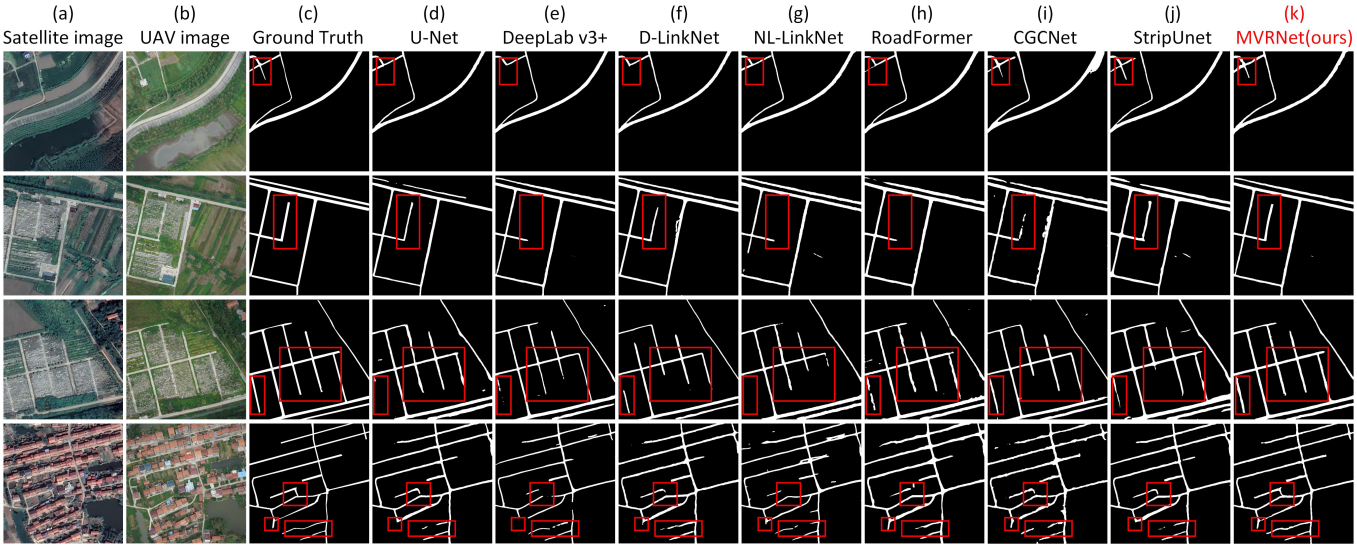


Fig. 6. Visual comparison of road extraction results on the SURoad dataset. The red boxes highlight regions with dense roads, complex structures, and high error susceptibility. Compared with the baseline methods, MVRNet demonstrates better road connectivity and structural integrity.

TABLE I
EXPERIMENTAL RESULTS OF ALL METHODS ON THE SUROAD DATASET

| Method | P (%) | R (%) | F1 (%) | IoU (%) |
|---|---|---|---|---|
| U-Net | 76.56 | 75.81 | 76.18 | 61.53 |
| DeepLab v3+ | 73.64 | 80.80 | 77.05 | 62.67 |
| D-LinkNet | 73.04 | 82.35 | 77.42 | 63.15 |
| NL-LinkNet | 76.82 | 76.55 | 76.69 | 62.19 |
| RoadFormer | 80.02 | 77.79 | 78.89 | 65.14 |
| CGCNet | 80.36 | 80.08 | 80.22 | 66.97 |
| StripUnet | 82.49 | 82.09 | 82.29 | 69.91 |
| **MVRNet (ours)** | **83.27** | **83.06** | **83.17** | **71.18** |

terms of IoU, which is a critical metric for segmentation quality, MVRNet achieves 71.18%, surpassing the best baseline, StripUNet (69.91%), by approximately 1.3%. This gain clearly shows the effectiveness of cross view fusion in overcoming the limits of single view feature representation. At the same time, the joint improvement in Precision and Recall reveals the internal cooperation within the model. By aligning shallow details and injecting deep topological information, the model can not only accurately detect narrow road branches but also recover occluded road segments in complex backgrounds. This balanced performance demonstrates that MVRNet has strong structural completeness and robustness in complex scenarios.

Fig. 6 presents a visual comparison in typical scenarios. As observed, in challenging regions like dense roads or partial occlusions, other methods generally exhibit varying degrees of road fragmentation or structural disconnection. In contrast, MVRNet generates prediction results that are more continuous, complete, and highly consistent with the ground truth. This indicates that the satellite structural priors introduced by the model play a decisive role in resolving road network fragmentation. This enables the model to accurately reconstruct physically consistent road topological structures in occluded regions where visual cues are lacking.

### C. Ablation Study

To validate the effectiveness of key modules in MVRNet, we conducted a series of ablation studies on the SURoad test set. The results are presented in Table II.

TABLE II
PERFORMANCE RESULTS OF KEY MODULES ABLATION

| Method | Baseline | CDPM | TAMM | DGRM | F1 (%) | IoU (%) |
|---|---|---|---|---|---|---|
| Method1 | ✓ | – | – | – | 68.99 | 52.66 |
| Method2 | ✓ | ✓ | – | ✓ | 82.57 | 70.32 |
| Method3 | ✓ | – | ✓ | ✓ | 82.76 | 70.59 |
| Method4 | ✓ | ✓ | ✓ | – | 80.71 | 67.66 |
| **MVRNet** | ✓ | ✓ | ✓ | ✓ | **83.17** | **71.18** |

Analyzing the results in Table II, Method 1, using only simple fusion, achieves an IoU of only 52.66%, indicating that simple concatenation fails to fully utilize multi view information. Method 2 significantly boosts the IoU by 17.66% (to 70.32%), demonstrating the effectiveness of CDPM in enhancing thin road details by aligning features with the DGRM restoration mechanism. Method 3 further enhances the performance, indicating that TAMM effectively utilizes structural priors for topological guidance. In contrast, results from Method 4 show that without DGRM, the IoU drops to 67.66%, suggesting that the lack of dynamic balancing leads to instability in complex areas. In comparison, the complete MVRNet achieves peak performance, strongly validating the functional complementarity and the layered collaborative modeling design.

### D. Effect of Cross-View Input Configuration

To verify the necessity of introducing satellite views and the core value of heterogeneous complementary information, we conducted an ablation study on input configurations.

TABLE III
PERFORMANCE RESULTS OF DIFFERENT INPUT CONFIGURATIONS

| Method | Input | P (%) | R (%) | F1 (%) | IoU (%) |
|---|---|---|---|---|---|
| MVRNet | UAV–UAV | 76.81 | 77.82 | 77.32 | 63.02 |
| **MVRNet** | **SAT–UAV** | **83.27** | **83.06** | **83.17** | **71.18** |

The results in Table III validate the effectiveness of the method from two aspects. First, the significant performance improvement under SAT-UAV input confirms the high complementarity between satellite macrostructures and UAV local details, establishing the necessity of multi-view inputs. Second, the performance leap brought by cross-view inputs indicates that MVRNet has a strong capability of heterogeneous information fusion, effectively converting this complementary advantage into practical improvements in segmentation accuracy.

## V. CONCLUSION

This paper proposes a multi-view fusion network, MVRNet, and constructs the SURoad dataset which serves as a benchmark for cross-view road extraction. Research demonstrates that the CDPM effectively rectifies geometric misalignments caused by view disparities, significantly enhancing boundary precision in shallow features. The TAMM successfully injects satellite macroscopic priors into deep layers, mitigating road network fragmentation caused by occlusions. Meanwhile, the DGRM achieves an optimal balance between topological consistency and local detail recovery through a dynamic gating mechanism. Overall, MVRNet not only fully exploits the complementary advantages of satellite and UAV images but also achieves efficient feature fusion through a layered collaborative modeling strategy, strongly validating its robustness and effectiveness in handling cross-view and heavily occluded scenarios.

Future work will focus on expanding the scale and diversity of the SURoad dataset and exploring more efficient cross-view feature interaction mechanisms to further exploit complementary information, thereby improving the model's generalization ability and robustness across diverse geographic scenarios.

## REFERENCES


[1] R. Liu, J. Wu, W. Lu, Q. Miao, H. Zhang, X. Liu, Z. Lu, and L. Li, "A review of deep learning-based methods for road extraction from high-resolution remote sensing images," *Remote Sensing*, vol. 16, no. 12, 2024.

[2] J. Chen, L. Yang, H. Wang, J. Zhu, G. Sun, X. Dai, M. Deng, and Y. Shi, "Road extraction from high-resolution remote sensing images via local and global context reasoning," *Remote Sensing*, vol. 15, no. 17, p. 4177, 2023.

[3] L. Zhao, J. Zhang, X. Meng, W. Zhou, Z. Zhang, and C. Peng, "Road extraction method of remote sensing image based on deformable attention transformer," *Symmetry*, vol. 16, no. 4, p. 468, 2024.

[4] X. Zhu, X. Huang, W. Cao, X. Yang, Y. Zhou, and S. Wang, "Road extraction from remote sensing imagery with spatial attention based on swin transformer," *Remote Sensing*, vol. 16, no. 7, p. 1183, 2024.

[5] Y. Wang, L. Tong, S. Luo, F. Xiao, and J. Yang, "A multiscale and multidirection feature fusion network for road detection from satellite imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–18, 2024.

[6] Y. Lin, L. Wan, H. Zhang, S. Wei, P. Ma, Y. Li, and Z. Zhao, "Leveraging optical and sar data with a uu-net for large-scale road extraction," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102498, 2021.

[7] H. Luo, Z. Wang, B. Du, and Y. Dong, "A deep cross-modal fusion network for road extraction with high-resolution imagery and lidar data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[8] T. Sun, Z. Di, P. Che, C. Liu, and Y. Wang, "Leveraging crowdsourced gps data for road extraction from aerial imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7509–7518.

[9] A. Batra, S. Singh, G. Pang, S. Basu, C. Jawahar, and M. Paluri, "Improved road connectivity by joint learning of orientation and segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 385–10 393.

[10] W. Boonpook, Y. Tan, B. Bai, and B. Xu, "Road extraction from uav images using a deep resdclnet architecture," *Canadian Journal of Remote Sensing*, vol. 47, no. 3, pp. 450–464, 2021.

[11] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *Proceedings of the 28th ACM international conference on Multimedia*, 2020, pp. 1395–1403.

[12] K. Regmi and A. Borji, "Cross-view image synthesis using conditional gans," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3501–3510.

[13] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] W. Xu, Y. Yao, J. Cao, Z. Wei, C. Liu, J. Wang, and M. Peng, "Uav-visloc: A large-scale dataset for uav visual localization," *arXiv preprint arXiv:2405.11936*, 2024.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[18] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 192–1924.

[19] Y. Wang, J. Seo, and T. Jeon, "Nl-linknet: Toward lighter but more accurate road extraction with nonlocal operations," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[20] X. Liu, Z. Wang, J. Wan, J. Zhang, Y. Xi, R. Liu, and Q. Miao, "Roadformer: Road extraction using a swin transformer combined with a spatial and channel separable convolution," *Remote Sensing*, vol. 15, no. 4, p. 1049, 2023.

[21] X. Ma, X. Zhang, D. Zhou, and Z. Chen, "Stripunet: A method for dense road extraction from remote sensing images," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 11, pp. 7097–7109, 2024.

[22] P. Liu, X. Gao, C. Shi, Y. Lu, L. Bai, Y. Fan, Y. Xing, and Y. Qian, "Cgcnet: Road extraction from remote sensing image with compact global context-aware," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–12, 2025.