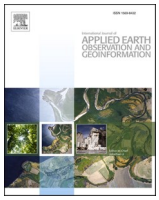
International Journal of Applied Earth Observation and Geoinformation 124 (2023) 103522

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Building and road detection from remote sensing images based on weights adaptive multi-teacher collaborative distillation using a fused knowledge

Ziyi Chen [a], Liai Deng [a], Jing Gou [a], Cheng Wang [b], Jonathan Li [c], Dilong Li [a,*]

[a] Department of Computer Science and Technology, Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Xiamen Key Laboratory of Data Security and Blockchain Technology, Huaqiao University, 668, Jimei Road, Xiamen, FJ 361021, China
[b] School of Informatics, Xiamen University, Xiamen, FJ 361005, China
[c] Department of Geography and Environmental Management and Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

ARTICLE INFO

Keywords:
Knowledge distillation
Remote sensing
Building extraction
Road extraction

ABSTRACT

Knowledge distillation is one effective approach to compress deep learning models. However, the current distillation methods are relatively monotonous. There are still rare studies about the combination of distillation strategies using multiple types of knowledge and employing multiple teacher models. Besides, how to optimize the weights among different teacher models is still an open problem. To address these issues, this paper proposes a novel approach for knowledge distillation, which effectively enhances the robustness of the distilled student model by a weights adaptive multi-teacher collaborative distillation. Moreover, the proposed method utilizes feature knowledge exchange guidance between teacher networks to transfer more comprehensive feature knowledge to the student model, which further improves the learning capability of hidden layers' details. The extensive experimental results demonstrate that the proposed method achieves state-of-the-art performance on Massachusetts Roads Dataset, LRSNY Roads Dataset, and WHU Building Dataset. Specifically, under the guidance of the first ensemble of teacher networks, we obtained IoU scores of 47.33%, 78.15%, and 80.71%, respectively. Under the guidance of the second ensemble of teacher networks, we obtained IoU scores of 48.56%, 79.51%, and 81.35%, respectively.

1. Introduction

With the advancement of scientific and technological capabilities, significant advancements have been made by neural networks in several fields, including categorization of images(Wang et al., 2021a), object detection(Li et al., 2020), identification of faces (Deng and Guo, 2018), image semantic segmentation (Chen et al., 2018b), and image retrieval (Yan et al., 2021). Image semantic segmentation has significant practical implications for remote sensing image processing. The remote sensing imagery serves as one of the primary tools for observing and monitoring the Earth's surface. In this context, the extraction of information (e.g. buildings and roads) from remote sensing images holds immense practical significance. This information not only plays a crucial role in urban planning (Schrotter and Hürzeler, 2020) and population estimation (Chen et al., 2021) but also exerts far-reaching impacts across various domains, including but not limited to traffic management and safety, natural disaster and resource management, as well as infrastructure planning. Researchers have conducted extensive research on remote sensing image extraction, yielding substantial research achievements. Presently, deep learning models serve as the primary foundation for extraction algorithms in the domain of remote sensing images. Nevertheless, these neural network models tend to possess a substantial number of parameters, ranging from several million to billions in magnitude, making them computationally intensive and reliant on high-performance server backends, rendering them unsuitable for direct application on satellite computing devices and edge computing devices. Knowledge distillation, which trains lightweight student models guided by complex models, enables the retention of student models with high accuracy and robustness. It is currently the mainstream approach for model lightweighting (Hinton et al., 2015).

However, methods of knowledge distillation are still the subject of exploratory study, with relatively limited diversity in distillation

* Corresponding author.
E-mail addresses: chenziyihq@hqu.edu.cn (Z. Chen), liaideng07@163.com (L. Deng), goujin@hqu.edu.cn (J. Gou), cwang@xmu.edu.cn (C. Wang), junli@uwaterloo.ca (J. Li), scholar.dll@hqu.edu.cn (D. Li).

https://doi.org/10.1016/j.jag.2023.103522
Received 30 June 2023; Received in revised form 19 September 2023; Accepted 13 October 2023
Available online 24 October 2023

techniques and insufficient exploration of distillation using multiple types of knowledge and multiple teacher types. Additionally, optimizing the weights among different teacher models is also an important topic that requires further research.

To release those issues, we put forward a knowledge distillation strategy that uses multiple teacher models and multiple types of knowledge. Specifically, this distillation method analyzes and integrates knowledge from different teacher networks, transferring three distinct types of knowledge to the student network based on different distillation losses, resulting in a significant improvement in its performance. When combining different teacher models, a weight adaptive evaluation module is also proposed to optimize the weights of the different teacher models. Furthermore, within the domain of remote sensing image extraction tasks, we employ a multi-task learning strategy to successfully combine the loss functions of several tasks, which improves the performance of the model.

The paper makes the following key contributions:

(1) We put forward a knowledge distillation strategy based on feature-level fusion among multiple teachers. This strategy involves mutual learning and fusion of feature knowledge between teacher networks, followed by guiding the student network in learning feature layers. This approach effectively enhances the robustness of distillation against erroneous guidance.
(2) Building upon the feature-level fusion between teachers, this paper introduces an innovative strategy for multi-teacher distillation that integrates multiple types of knowledge. During the fusion of multiple knowledge types, we combine the knowledge from feature-level fusion among teachers, relation-based knowledge, and response-based knowledge for distillation.
(3) This paper introduces an effective strategy for multi-knowledge fusion and multi-teacher collaborative distillation. Furthermore, an adaptive weight assessment algorithm is integrated during the multi-teacher distillation stage, ensuring the precision of teacher network guidance throughout the training process.

2. Related work

2.1. Knowledge distillation

Knowledge distillation is an approach aimed at compressing and accelerating models. It enables the effective improvement of the performance of lightweight student model's performance under knowledgeable teacher model's direction, thus achieving model compression. The concept of knowledge distillation, also known as 'dark knowledge extraction', was first introduced by Hinton in 2014 (Hinton et al., 2015). It entails incorporating the teacher network's soft objectives into the aggregate loss function, which directs the compact model's training and promotes knowledge transfer. In recent years, an increasing number of researchers and scholars have recognized the promising performance and significant application value of knowledge distillation in achieving model lightweighting. Currently, knowledge categories in knowledge distillation are typically divided into three major groups: knowledge based on responses, knowledge based on features, and knowledge based on relation.

To address the issue of inconsistent sample and label sizes during neural network training, Bagherinezhad et al. (2018) enhanced the quality of labels through iterative guided training, thereby further improving the model's generalization ability. Yim et al. (2017) found that hint training can represent the training process and achieve knowledge transfer, fast convergence, and transfer learning by fine-tuning the student network on the target task dataset. Heo et al. (2018) employed adversarial attack strategies to convert benchmark class samples into target class samples and used the generated adversarial examples to steer the development of the student network. This approach significantly enhances the student network's ability to recognize decision boundaries. Considering the neglect of logit distillation in current knowledge distillation methods, Zhao et al. (2022) introduced the Decoupled Knowledge Distillation (DKD) approach, which efficiently combines the benefits of target class knowledge distillation with non-target class knowledge distillation, leading to a more flexible and effective knowledge distillation process, thereby further improving the model's generalization ability. Furlanello et al. (2018) combined the distilled student models with the teacher model to obtain better generalization performance on test data.

To provide guidance for training deeper and more compact student networks, Romero et al. (2014) developed a better knowledge distillation technique. This technique utilized the middle-tier feature map in the teacher network to provide guidance within the corresponding guidance layer of the student network. In their exploration of knowledge distillation, Zagoruyko and Komodakis (2016), adjusted the features of attention regions from the perspective of attention mechanisms. This differs from the carefully designed knowledge for guiding student network training in known knowledge distillation tasks. To address the reliance on a large number of pre-trained estimators in existing pose distillation works, Li et al. (2021b) introduced an online knowledge extraction framework for human posture. This framework improves distillation efficiency by extracting human pose structure knowledge. Chen et al. (2022a) stressed the need of reducing the performance gap between instructor and student models in an effective manner, direct knowledge distillation methods are commonly employed. They recommended employing a pre-trained model's discriminative classifier as the teacher to guide student training. In knowledge distillation methods, different architectures may cause semantic information at the same spatial location to differ. To tackle this problem, Lin et al. (2022) proposed a full-space matching knowledge distillation method that maps each pixel of teacher features to all spatial positions of student features.

Park et al. (2019) effectively utilized associations between sample characteristics as knowledge and transferred this information to offer direction for the development of the compact network from the teacher model. Xu et al. (2020) introduced self-supervised learning as an additional task to complement knowledge extraction. Through this approach, the model demonstrates effectively discriminates between positive and negative samples, thereby enhancing its learning of task-specific knowledge and feature representation capabilities. This approach significantly enhances the model's classification performance. In contrast to methods that require pre-training a single teacher, in (Lin et al., 2017a; Lin et al., 2017b), multiple students are guided simultaneously, facilitating mutual learning among the student networks. Ji et al. (2021) proposed Feature Refinement with Self-Knowledge Distillation (FRSKD), aiming to address the issue of losing local information due to data augmentation during large-scale teacher model training and self-knowledge distillation. FRSKD combines soft labels and feature map distillation techniques to achieve self-knowledge distillation more effectively. Yang et al. (2022) introduced Cross-Image Relationship Knowledge Distillation (CIRKD), which aims to address the problem of neglecting global semantic relationships between individual images and pixels in conventional knowledge distillation (KD) approaches used in semantic segmentation. This approach combines structured pixel-to-pixel and pixel-to-region relationships between entire images as distillation losses to improve the student network's ability to replicate the structured semantic relationships of the teacher network.

Yuan et al. (2020) introduced the use of label smoothing regularization to add a virtual teacher model to a knowledge distillation task. They further proposed the framework of Teacher-free Knowledge Distillation (Tf-KD), where the learning of regularized distributions, either on its own or manually designed, can optimize student models. Zhang et al. (2022) used a method known as using generative adversarial networks for image-to-image translation tasks. They addressed the challenge of producing high-quality, high-frequency information has been addressed using Wavelet Knowledge Distillation. Instead of directly extracting information from the generated samples, this

approach utilizes discrete wavelet transforms to decompose the image into various frequency bands to obtain high-frequency band information. Kang et al. (2021) recommended a conditional distillation framework for knowledge extraction to overcome the challenges faced by knowledge distillation in object detection. The framework incorporates a trainable conditional decoding module that retrieves information for each target instance based on queries. To tackle the challenge of self-supervised pretraining for small models, Bhat et al. (2021)proposed a knowledge distillation approach named Domain Guided online Knowledge distillation (DoGo) that enhances the performance of compact networks by using single-stage online knowledge distillation.

2.2. Object extraction from remote sensing images

Recent years have seen a tremendous advancement in remote sensing technology and a broad use of high-resolution satellite images. Significant advancements in the process of extracting things from remote sensing images have been made by object identification algorithms based on deep learning. Numerous cutting-edge approaches for extracting roads and buildings have been presented and are being continually improved by using the benefits of deep learning. Recent research has demonstrated the effectiveness of deep learning-based techniques in accurately extracting roads and buildings from remote sensing images.

2.2.1. Road extraction in remote sensing images

Road extraction holds significant importance in remote sensing image information extraction. Mnih and Hinton introduced a technique that employs Restricted Boltzmann Machines (RBMs) to identify road regions in high-quality aerial images (Mnih and Hinton, 2010), marking the initial attempt that deep learning techniques are combined with remote sensing images road extraction. Wu et al. (2021) proposed the Dense Global Residual Network (DGRN) to reduce the enhances contextual awareness. To improve the extraction of local and global data from remote sensing, Luo et al. (2022) introduced a Bidirectional Transformer Network (BDTNet) that utilizes a hybrid encoder-decoder architecture. Wang et al. (2022a) introduced Dual-Decoder-U-Net (DDU-Net), a deep learning model that has been improved with an enhanced deep neural network model, which in duties involving the extraction of roads from remote sensing images, the reliability and accuracy of small roads are improved when there are different sizes of roads. You et al. (2022) suggested the Foreground Mixture Improved to Weighted Dual-Network Cross Training (FMWDCT) method for semi-supervised road extraction. The challenge of sample imbalance in road extraction task is addressed by this method. Zhou et al. (2022b) introduced SOC-RoadNet, a weakly supervised road segmentation network, which learns road information from open-source road maps using structural and directional consistency principles, the extraction of high-quality, extensive roads possible. Chen et al. (2022b) presented an adversarial learning-based semi-weakly supervised approach for extracting road networks in remote sensing imagery, making full use of weak annotations in the dataset. Wang et al. (2021b) presented a method to enhance the accuracy and connectedness of road extraction in remote sensing images by using an Inception-Convolution Inherited Encoder-Decoder network. Zhou et al. (2022a) introduced a Segmentation Depthwise Separable Graph Convolutional Network (SGCN) that the precision of road extraction from high-resolution remote sensing images has witnessed significant improvement. Shi et al. (2014) proposed a road border recognition and ground point separation in remote sensing images are made possible by a method based on polar grids that makes use of trajectory data and feature filters. Hu et al. (2015) proposed a technique for detecting road boundaries that leverages Conditional Generative Adversarial Networks (CGANs) and converts point clouds into two-dimensional images.

2.2.2. Building extraction in remote sensing Images

Numerous fields, including urban planning (Schrotter and Hürzeler, 2020), population estimation (Chen et al., 2021), change detection (Chen et al., 2021), land use management (Chen et al., 2019), and other geographic and societal applications, can benefit from high-resolution remote sensing images. Currently, the success of deep learning in extracting buildings from remotely sensed images is mainly attributed to its ability to effectively capture and represent protrusions or prominent features of buildings. Wang et al. (2022b) simplified the training task for extracting architectural information using deep convolutional neural networks, a building feature prominence, global perception and cross-layer information fusion network (B-FGC-Net) including recommend spatial attention units and residual learning. To enhance the automated extraction of building boundaries, Zhou et al. (2022c) retain the morphological properties of extracted buildings, the problem of boundary optimization and fully segmented building extraction is created, together with a multi-scale context-aware network (BOMSC-Net). He and Jiang (2021) in the problem of maintaining the accuracy of boundary construction, embed boundary learning tasks in fully convolutional networks. Xia et al. (2021) semi-supervised learning is initially used to edge detection neural networks in the challenge of getting the roof border of structures in high-resolution remote sensing images. In reducing the dependency on labeled samples, this method co-trains the model by using smaller sample sets with labels and large batches of images without labels. Chen et al. (2022d) proposed an Encoder-Decoder network with Contour Guidance and Local Structure Awareness (CGSANet) to address the problem with current encoder-decoder architectures based on implicit features of extracting building forms from remote sensing images are not effectively used by FCN. Considering the complexity of building colors and textures in remote sensing images with high resolutions, Hosseinpour et al. (2022) introduced an end-to-end Cross-Modal Gated Fusion Network (CMGFNet), which extracts building from Very High-Resolution (VHR) remote sensing images and Digital Surface Model (DSM) data. Li et al. (2022) introduced an original Hierarchical Deconvolutional Network (HDNet) with a feature representation in convolutional neural networks during the extraction of remote sensing images is unstable and imprecise. This is addressed by the encoder-decoder structure. Given the extensive parameterization in current deep neural network-based approaches, which extracts building in remote sensing images, Chen et al. (2022c) proposed a Context Feature Enhancement Network (CFENet). Guo et al. (2021) presented a Scene-Driven Multi-Task Parallel Attention Convolutional Network (MTPA-Net) to solve the limitation that current building extraction methods based on convolutional neural networks cannot cover buildings in different scenes. Lei et al. (2022) propose selective non-local resUNeXt++ (snlrux++), which aims to improve the robustness of the semantic segmentation model for remote sensing image extraction tasks.

3. Methodology

In this section, we will introduce the proposed knowledge distillation strategy that combines multiple teachers, integrates multiple knowledge sources, and facilitates feature knowledge exchange among teacher networks. Then, we will present and discuss each key strategy employed in the distillation process. Finally, we will present a comprehensive explanation of the distillation loss employed in each strategy.

As shown in Fig. 1, our proposed strategy for distillation, which combines multiple types of knowledge and utilizes a collaborative ensemble of teacher networks, consists of three main components: knowledge distillation based on feature fusion among multiple teachers, collaborative distillation with multiple teachers, and relation-based fusion of multiple types of knowledge.

We utilize classical deep learning methods for remote sensing image extraction, such as the U-Net (Ronneberger et al., 2015) and DeepLabV3Plus (Chen et al., 2018a), to construct multiple teacher network

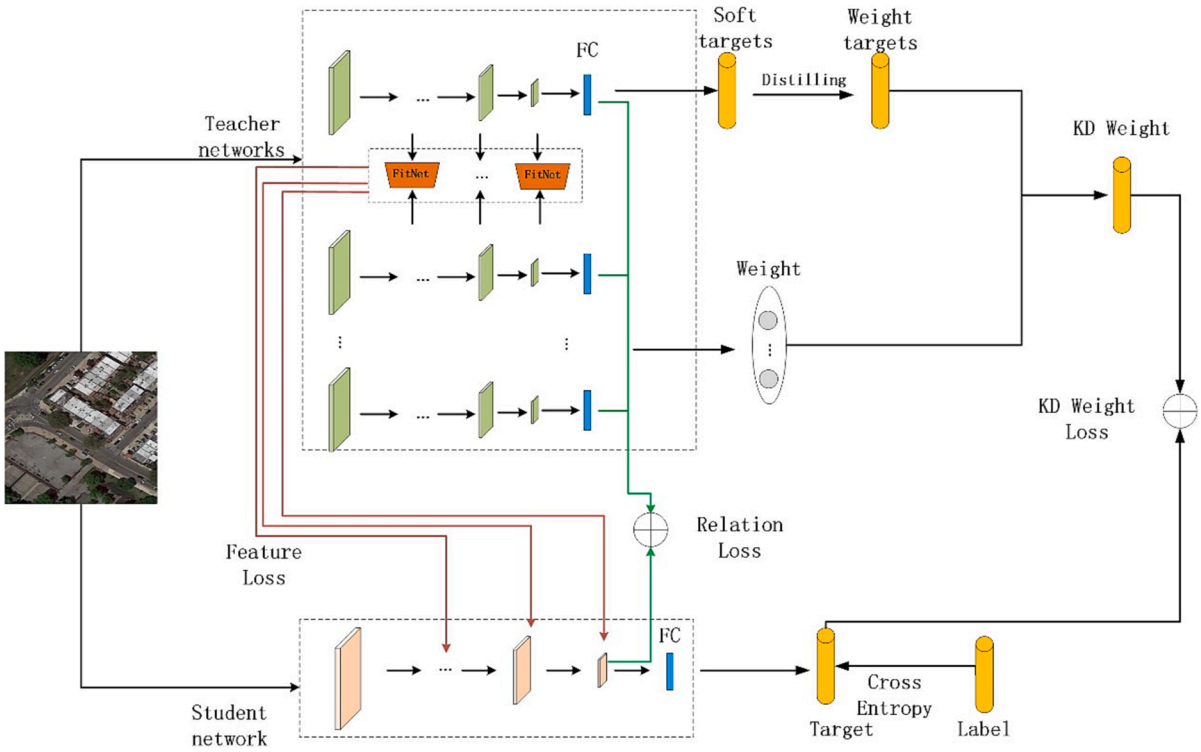


Fig. 1. Overview of our proposed distillation strategy that combines multiple knowledge sources and utilizes a collaborative ensemble of teacher networks.

models, each having independent feature representations and knowledge. In the knowledge distillation module based on feature fusion among multiple teachers, we first design a convolutional regression-based feature fusion module. After effective feature fusion, the trained student network's feature maps are guided. This is depicted by the solid red line in Fig. 1.

In the module on joint distillation among multiple teachers, we first employ a weight auto-evaluation module to assess the importance of different teacher models and determine the guiding weights for each teacher network. In the weight auto-evaluation module, the cross-entropy loss between the real label and the teacher network prediction is calculated as part of the sample-based weight assessment process. The inflexibility brought on by set teacher weights is successfully overcomed by using this to calculate the distribution of teacher model weights for each sample.

Euclidean distance is used in the relation-based multi-knowledge fusion distillation module to gauge the relationship correlation between several targets. First, by computing the Euclidean distance between the teacher networks' output characteristics, we obtain a representation of the relationships between targets. Then, we leverage this relationship representation to enhance the student model's comprehension of the interplay and correlations between the targets, thereby enhancing the effectiveness of knowledge transfer during the distillation process. This is depicted by the solid green line in Fig. 1.

3.1. *Knowledge distillation based on multi-teacher feature fusion*

In the realm of knowledge distillation, we believe that learning from intermediate feature layers contributes to improving the performance of lightweight networks. Therefore, we extend the distillation method to intermediate layers to extract more useful information from the teacher networks. Firstly, we design a feature fusion module based on convolutional regression, which effectively merges the features before guiding the training of the student network's feature layers. As shown in Fig. 2, initially, we apply padding to the feature maps extracted from the teacher networks, resulting in the creation of novel feature maps.

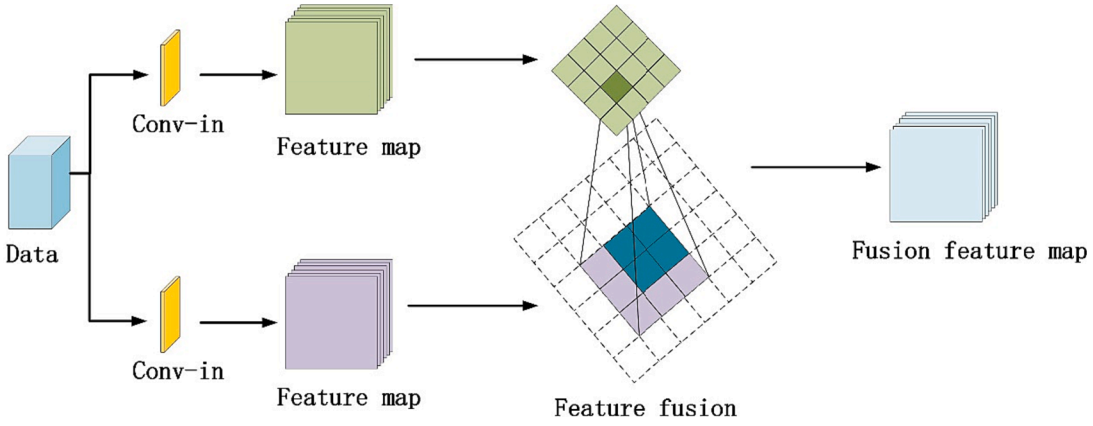


Fig. 2. Feature-level fusion among multiple teachers.

Subsequently, we utilize a 3 × 3 convolutional kernel on these newly formed feature maps, effectively achieving the fusion of the two feature maps.

Then, we measure the alignment between the teacher networks' feature layers by comparing their representations at the intermediate level using mean squared error loss. Finally, we combine the classification loss and mean squared error loss to train the student network. The mean squared error loss formulation for the teacher's intermediate feature layers is as follows:

$$L_{fearure} = \left(t_{ji} - t_{ki}\right)^2, \tag{1}$$

where t_{ji} denotes the i-th feature of the j-th teacher model.

3.2. Multi-teacher collaborative knowledge distillation

In order to increase the trained student network's resilience and accuracy, as well as its capacity to adjust to ineffective instructor models. Hence, we have devised a distillation strategy that integrates a collaborative multi-teacher network approach, engaging multiple teachers in guiding the training of a lightweight network. In the process of allocating guidance weights to the teacher networks, inspired by Zhang et al. (2021), we have developed an adaptive weight allocation algorithm to efficiently distribute the guidance weights of the teacher networks.We can determine the estimated weight distribution by calculating the cross-entropy loss between the predictions made by the teacher network and the labels assigned to the data. The final model predictions in knowledge distillation are obtained using the softmax function with temperature T, given by $\sigma(z) = \frac{\exp(Z^c/T)}{\sum_j \exp(Z^c/T)}$, where $z = [z^1, z^2, \ldots, z^c]$ represents the output logits and C is the number of classes. The distillation loss is shown as follows:

$$L^k_{CEkd} = -\sum_{c=1}^{C} y^c \log\left(\sigma\left(z^c_{T_k}\right)\right). \tag{2}$$

$$w^k_{KD} = \frac{1}{K-1}\left(1 - \frac{\exp\left(L^k_{CEkd}\right)}{\sum_j \exp\left(L^k_{CEkd}\right)}\right). \tag{3}$$

Here T_k represents the k-th teacher, and as L^k_{CEkd} decreases, the corresponding w^k_{KD} increases. Then, the overall predictions of the teachers are combined with the computed weights:

$$L_{KD} = -\sum_{k=1}^{K} w^k_{KD} \sum_{c=1}^{C} z^c_{T_k} \log\left(\sigma\left(z^c_S\right)\right). \tag{4}$$

3.3. Relation-based multi-knowledge fusion distillation

To further explore useful information from the teacher networks, we introduce a knowledge fusion approach based on relations for distillation. As shown in Fig. 3, We compute pairwise distances between the multiple outputs of both teacher and student networks within each batch, ultimately forming a relationship-based structural output of size batch × batch.

In this section, we utilize Euclidean distance to measure the relationship correlation between different targets. When transferring relation-based knowledge, we employ the Smooth L1 loss function to mitigate the impact of large errors, as it imposes a smaller penalty on such errors, reducing the impact of outliers and facilitating better knowledge transfer of relationships. By incorporating relation-based knowledge distillation, we leverage the inter-object relation information in semantic segmentation tasks to improve the student model's performance. The loss expression is as follows:

$$L_{relation} = l\left(\frac{1}{\phi(t)}\|t_i - t_j\|_2, \frac{1}{\phi(t)}\|s_i - s_j\|_2\right). \tag{5}$$

$$\phi(x) = \|x_i - x_j\|_2. \tag{6}$$

Here t_i, t_j, s_i, s_j represent the i-th feature of the teacher and student models, where $i \neq j$. $\phi(x)$ is the normalization factor for distance, and $l(\bullet)$ represents L_1 smooth loss function.

3.4. The overall loss function

In addition to the three aforementioned losses, we also calculate the regularized cross-entropy loss of the ground truth labels. The loss function is as follows:

$$L_{CE} = -[y log\hat{y} + (1-y)\log(1-\hat{y})]. \tag{7}$$

The overall loss function of our knowledge distillation method is as follows:

$$L = L_{CE} + \alpha L_{feature} + \beta L_{relation} + \gamma l_{KD}, \tag{8}$$

where α, β and γ are hyperparameters used to control the balance between the impact of knowledge distillation and the standard cross-entropy loss.

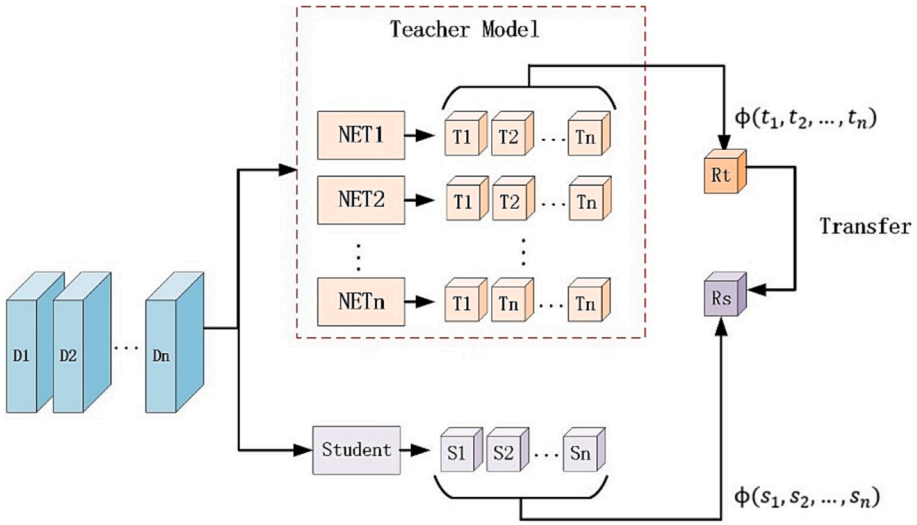


Fig. 3. Feature-level fusion among multiple teachers.

4. Experiments

4.1. Experimental preparation

4.1.1. Datasets

On the Massachusetts Roads dataset, the LRSNY dataset, and the WHU Building dataset, we carried out comprehensive trials to confirm the efficacy and reliability of our proposed strategy.

(1) The Massachusetts Roads dataset comprises 1,171 aerial images captured in Massachusetts, encompassing urban, suburban, and rural areas. The dataset includes optical remote sensing images, each containing an area spanning 2.25 square kilometers and measuring 1500 × 1500 pixels. 1,108 images from the training set, 14 images from the validation set, and 49 images from the testing set make up the dataset. In our experiments, we divided each image into small patches of 256 × 256 pixels. Fig. 4 displays several sample cut images of this size used in our experiments. Each image has a corresponding ground truth that is represented as a binary image, where road areas are represented in white (255, 255, 255), and background areas are represented in black (0, 0, 0). The dataset can be obtained from the following website: *http://www.cs.toronto.edu/~vmnih/data/*.

(2) The LRSNY roads dataset contains images of the central part of New York City, captured at a resolution of 0.5 m. The original images are 37949 × 35341 pixels in size, and for user convenience and standardisation, they have been broken into smaller blocks of 1000 × 1000 pixels. 716, 220, and 432 images total from this dataset are used for training, validation, and testing. The images are also available in a 256 × 256 pixel format in the dataset. Fig. 5 shows several sample cut images of this size used in our experiments. The dataset can be obtained from the following website: *https://pan.baidu.com/s/1jkKPjLYeadRipLGzTNxLgA*.

(3) The WHU Building Dataset is a dataset released by the Remote Sensing Information Engineering Research Center at Wuhan University for building extraction and classification. The dataset consists of two subsets: WHU Building Dataset I and II. WHU Building Dataset I includes high-resolution remote sensing images, building masks, and building height information, for a total of 25 images. WHU Building Dataset II contains more data, with 20 scenes in total. Each scene includes remote sensing images with four bands, corresponding masks, and building height information. In our experiments, we divided each image into small patches of 256 × 256 pixels. Fig. 6 displays several sample cut images of this size used in our experiments.

4.1.2. Network architecture

In all experiments, our first set ensemble of teacher networks includes the U-net and DeepLabV3Plus networks with ResNet101 as the backbone. The network branches of U-Net consist of convolutional layers, pooling layers, and upsampling layers, with a convolutional kernel size of 3 × 3. The network branches of DeepLabV3Plus comprise a ResNet101 backbone, ASPP module, and decoder module, with a convolutional kernel size of 3 × 3. The parameter count of the U-net is 3.1×10^7, and the parameter count of DeepLabV3Plus is 5.9×10^7. The second set ensemble of teacher networks includes the CRAE-Net (Li et al., 2021a) and SGCN (Zhou et al., 2022a). The network branches of CRAE-Net include ResNet50 as the backbone feature extraction network, positional attention module, channel attenuation module, and residual blocks. The network branches of CRAE-Net consist of ResNet50 as the backbone feature extraction network, feature separation, graph construction, and decoder modules. The parameter count of the CRAE-Net is 4.9×10^7. The parameter count of the SGCN is 4.2×10^7. For the student network, we employ the lightweight network BiSeNetV2 with a parameter count of 3×10^6.

4.1.3. Evaluation metrics

We employ four widely-accepted assessment measures for road segmentation performance to fully assess the performance of the models: precision, recall, and F1 score. These metrics are computed as follows:

$$Precision = \frac{FP}{FP + TP}. \tag{9}$$

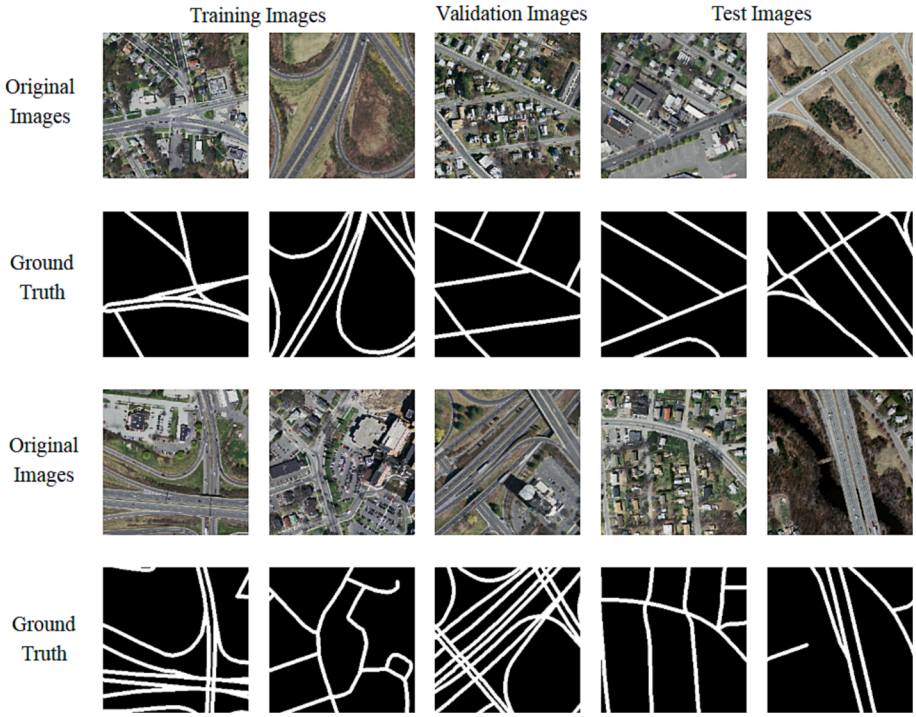


Fig. 4. Example of the Massachusetts Roads dataset.

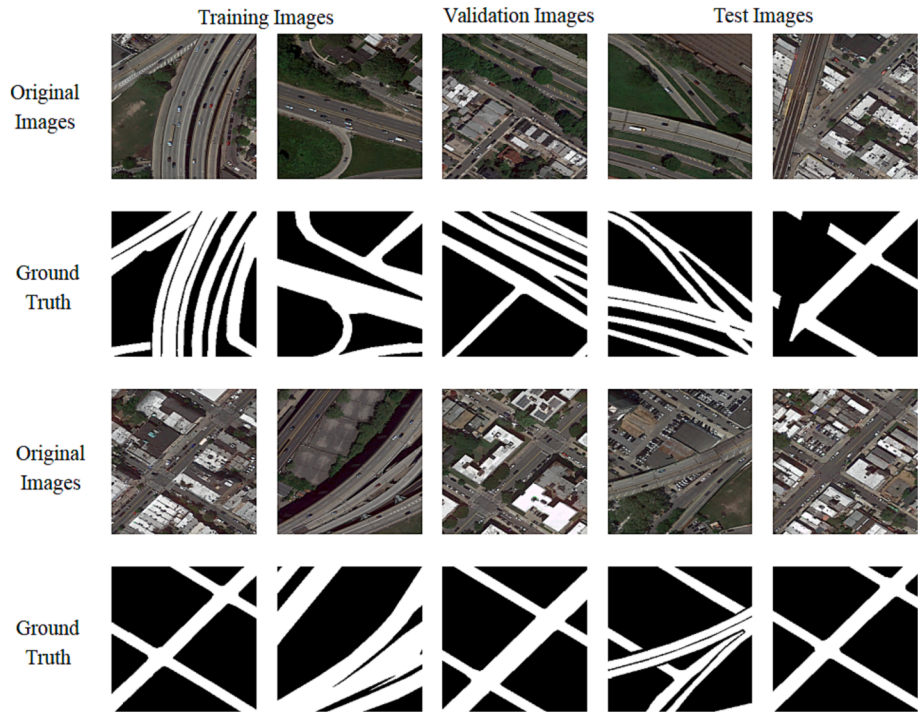


Fig. 5. Example of LRSNY Roads Dataset.

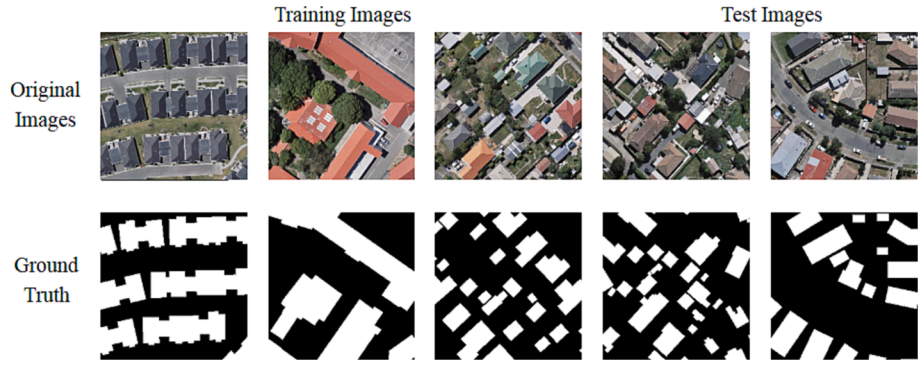


Fig. 6. Example of WHU Buildings Dataset.

$$Recall = \frac{FP}{FP + FN} \tag{10}$$

$$F1 - Score = \frac{2*precision*recall}{precision + recall}. \tag{11}$$

$$\text{IoU} = \frac{TP}{TP + FN + FP}. \tag{12}$$

Among these metrics, the classification of predicted bounding boxes is determined based on the IOU with a specified threshold (default is 0.5). If the IOU is equal to or greater than the threshold, it is considered a true positive (TP); otherwise, it is classified as a false positive (FP).

4.1.4. Training details

All experiments were optimized using the Adam optimizer with a batch size of 16, the learning rate for distillation was initially set at 0.0005, and a learning rate decay method was used to reduce it by doubling every 40 iterations. The starting learning rate for single-network training was set at 0.0001. The PyTorch framework was used to conduct the tests, and training on the building and road datasets took place across 200 iterations. The most accurate models developed via training were kept. The experimental setup included an Intel(R) Core (TM) i9-12900KF CPU @ 3.20 GHz, 128 GB of RAM, and 2 GPUs (NVIDIA GeForce RTX 3090) for training.

To analyze the training efficiencies of our proposal, we compared the training time and training effects among other distillation strategies and our proposed strategy. As shown in Table 1, compared to single knowledge distillation strategies, although our training time has increased, our IoU performance shows better results. Furthermore, when compared to the WKD distillation strategy, we have achieved better performance in both training time and IoU, demonstrating that our

Table 1
The comparison of training time and training effects among our strategy and other distillation strategies tested on the LRSNY road dataset.

Method	Time(hour)	IOU
KD (Hinton et al., 2015)	0.46	0.7615
Finets (Romero et al., 2014)	0.68	0.7684
Relation (Park et al., 2019)	1.5	0.7701
WKD (Zhang et al., 2021)	2.6	0.7757
DKD (Zhao et al., 2022)	0.53	0.7698
Ours	2.1	0.7815

distillation strategy achieves a better compromise between model performance and training efficiency.

4.1.5. Compared distillation methods

We compared our proposed distillation strategy with several other distillation strategies: KD (Hinton et al., 2015), FitNets (Romero et al., 2014), Relation (Park et al., 2019), WKD (Zhang et al., 2021), DKD (Zhao et al., 2022). All methods were evaluated using the first set of teacher network.

To assess the effectiveness of our approach within the current landscape of building and road extraction research, we established a secondary ensemble of teacher networks, employing CRAE-Net and SGCN networks as our teacher models.

4.2. Results

On three remote sensing image datasets: the Massachusetts Roads Dataset, the LRSNY Roads Dataset, and the WHU Building Dataset, we evaluated and compared the performance of the proposed distillation approach with different distillation techniques in this section in terms of object extraction.

4.2.1. The results on Massachusetts Roads Dataset

In Table 2, we compared our proposed knowledge distillation strategy with other distillation methods. Bold numbers indicate the best results in each category. The student network is designated as S, while the instructor network is designated as T. The denotations are the same for the following tables. The last column of the table represents the network parameter count. The first row represents the outcomes of separately training the Unet teacher network, while the second row represents the outcomes of independently training the DeepLabV3Plus teacher network. The third row displays the results of independently training the student network BiSeNetV2. The fourth row to the fifth rows illustrate the outcomes of training the student network while being guided by response-based knowledge, feature-based knowledge, and relation-based knowledge, respectively. When directed and controlled by teacher networks, all distillation techniques successfully improve the compact network's performance. However, our proposed strategy outperform other approaches in terms of all evaluation metrics. The precision, recall, IoU, and F1 score achieved by our distillation strategy are 68.50%, 60.62%, 47.33%, and 64.25%, respectively. Compared to training the student network independently, the performance is improved by 1.65% in precision, 3.26% in recall, 2.35% in IoU score, and 2.09% in F1 score. Additionally, compared to other cutting-edge distillation technologies, our proposed method is superior, achieving a precision gain of 0.59% and an IoU gain of 0.53%. These results demonstrate that our distillation strategy, compared to other proposed strategies, can effectively guide the student network to achieve better segmentation performance.

In Table 3, we established a secondary ensemble of teacher networks, employing CRAE-Net and SGCN networks as our teacher models. The first row represents the outcomes of separately training the CRAE-Net teacher network, while the second row represents the outcomes of independently training the SGCN teacher network. As can be observed, with the second ensemble of teacher networks guiding the way, our proposed distillation strategy consistently delivers significant performance enhancements for the student network. The precision, recall, IoU, and F1 score achieved by our distillation strategy are 69.92%, 61.34%, 50.56%, and 66.49%, respectively. Compared to training the student network independently, the performance is improved by 3.07% in precision, 3.98% in recall, 3.58% in IoU score, and 3.33% in F1 score. Compared to the distillation performance indicated by the first group of teacher networks, the performance is improved by 1.42% in precision, 0.62% in recall, 1.23% in IoU score, and 1.24% in F1 score.

Table 2
Performance comparison with other distillation strategies on the Massachusetts roads dataset.

Method	Precision	Recall	IOU	F1
Unet(T)	0.7600	0.7134	0.6055	0.7048
Deeplabv3Plus(T)	0.7797	0.7019	0.6134	0.7277
BiSeNetV2(S)	0.6685	0.5736	0.4498	0.6216
+KD (Hinton et al., 2015)	0.6717	0.6049	0.4669	0.6386
+Finets (Romero et al., 2014)	0.6799	0.6003	0.4680	0.6376
+Relation (Park et al., 2019)	0.6699	0.5856	0.4575	0.6327
+WKD (Zhang et al., 2021)	0.6785	0.6050	0.4678	0.6401
+DKD (Zhao et al., 2022)	0.6791	**0.6082**	0.4666	0.6408
+Ours	0.6850	0.6062	0.4733	0.6425

Table 3
Distillation performance of SOTA building and road extraction on the Massachusetts roads dataset.

Method	Precision	Recall	IOU	F1
CRAE-Net (T)	0.7995	0.7633	0.6522	0.7812
SGCN (T)	0.7838	0.7376	0.6301	0.7790
BiSeNetV2(S)	0.6685	0.5736	0.4498	0.6216
+Ours	0.6992	0.6134	0.4856	0.6549

4.2.2. Results on the LRSNY Roads Dataset

In Table 4, we compared our proposed distillation method with various distillation techniques, evaluating them on the LRSNY roads dataset. Our proposed method consistently demonstrated superior performance. The precision, recall, IoU, and F1 score achieved by our distillation strategy are 88.85%, 86.85%, 78.15%, and 87.74%, respectively. These results represent improvements of 1.6%, 1.29%, 2.94%, and 1.85%, respectively, compared to the individually trained student network. Additionally, our distillation method outperformed other top-performing distillation strategies with a 0.58% IoU gain and a 0.43% F1 score gain. These results demonstrate that our proposed distillation strategy can guide students to achieve better segmentation performance compared to other proposed distillation strategies.

In Table 5, we established a secondary ensemble of teacher networks, employing CRAE-Net and SGCN networks as our teacher models. Clearly, with the second ensemble of teacher networks guiding the way, our proposed distillation strategy consistently delivers significant performance enhancements for the student network. The precision, recall, IoU, and F1 score achieved by our distillation strategy are 89.26%, 88.34%, 79.51%, and 88.62%, respectively. Compared to training the student network independently, the performance is improved by 2.01% in precision, 2.78% in recall, 4.3% in IoU score, and 2.73% in F1 score. Compared to the distillation performance indicated by the first group of teacher networks, the performance is improved by 0.41% in precision, 1.49% in recall, 1.36% in IoU score, and 0.88% in F1 score.

4.2.3. Results on the WHU building dataset

In Table 6, we contrast various distillation techniques tested on the WHU building dataset with our proposed knowledge distillation strategy. The strategies we propose consistently show optimal performance.

Table 4
Performance comparison with other distillation strategies on the LRSNY roads dataset.

Method	Precision	Recall	IOU	F1
Unet(T)	0.8957	0.8764	0.7953	0.8860
Deeplabv3Plus(T)	0.8998	0.9131	0.8289	0.9064
BiSeNetV2(S)	0.8725	0.8556	0.7521	0.8589
+KD (Hinton et al., 2015)	0.8775	0.8653	0.7615	0.8710
+Finets (Romero et al., 2014)	0.8744	0.8637	0.7684	0.8690
+Relation (Park et al., 2019)	0.8800	0.8587	0.7701	0.8706
+WKD (Zhang et al., 2021)	0.8817	0.8649	0.7757	0.8724
+DKD (Zhao et al., 2022)	0.8857	**0.8695**	0.7698	0.8731
+Ours	0.8885	0.8685	0.7815	0.8774

Table 5
Distillation performance of SOTA building and road extraction on the LRSNY roads dataset.

Method	Precision	Recall	IOU	F1
CRAE-Net (T)	0.9343	0.9321	0.8364	0.9200
SGCN (T)	0.9222	0.9265	0.8489	0.9301
BiSeNetV2(S)	0.8725	0.8556	0.7521	0.8589
+Ours	0.8926	0.8834	0.7951	0.8862

Table 6
Performance comparison with other distillation strategies on the WHU Building Dataset.

Method	Precision	Recall	IOU	F1
Unet(T)	0.9180	0.9334	0.8615	0.9250
Deeplabv3Plus(T)	0.9231	0.9302	0.8633	0.9266
BiSeNetV2(S)	0.8582	0.8882	0.7726	0.8693
+KD (Hinton et al., 2015)	0.8812	0.8923	0.7901	0.8875
+Finets (Romero et al., 2014)	0.8813	0.8990	0.7952	0.8921
+Relation (Park et al., 2019)	0.8752	0.8943	0.7858	0.8837
+WKD (Zhang et al., 2021)	**0.8849**	0.8957	0.8022	0.8903
+DKD (Zhao et al., 2022)	0.8810	0.8996	0.8011	0.8902
+Ours	0.8847	0.9020	0.8071	0.8933

The precision, recall, IoU, and F1 score achieved by our distillation strategy are 88.47%, 90.20%, 80.71%, and 89.33%, respectively. Compared to the individually trained student network, our method improves the performance by 2.65% in precision, 1.38% in recall, 3.45% in IoU score, and 2.4% in F1 score. Furthermore, our distillation method better than other cutting-edge tactics with a 0.49% IoU gain and a 0.24% recall gain. These results demonstrate that our proposed distillation strategy can guide students to achieve better segmentation performance compared to other proposed distillation strategies.

In Table 7, we established a secondary ensemble of teacher networks, employing CRAE-Net and SGCN networks as our teacher models. As can be observed, with the second ensemble of teacher networks guiding the way, our proposed distillation strategy consistently delivers significant performance enhancements for the student network. The precision, recall, IoU, and F1 score achieved by our distillation strategy are 90.02%, 90.98%, 81.35%, and 90.28%, respectively. Compared to training the student network independently, the performance is improved by 4.2% in precision, 2.16% in recall, 4.09% in IoU score, and 3.35% in F1 score.Compared to the distillation performance indicated by the first group of teacher networks, the performance is improved by 1.55% in precision, 0.78% in recall, 0.64% in IoU score, and 0.95% in F1 score.

According to Table 2, Table 4, and Table 6, in the task of extracting objects from remotely sensed images, our proposed knowledge distillation strategy performed very well, achieving the highest IoU and F1 scores on the Massachusetts Road Dataset, LRSNY Road Dataset, and WHU Building Dataset. On the Massachusetts Roads Dataset, our method outperforms other strategies in all four performance metrics, with improvements of 0.51%, 0.59%, 0.53%, and 0.49% compared to the second-best performing Finets distillation strategy, showcasing the effectiveness of our method in accurately extracting roads from remote sensing images. On the LRSNY Roads Dataset, in terms of IoU score, our approach performs better than the other three distillation techniques by 2%, 1.31%, and 1.14%. According to Table 6, our proposed distillation strategy demonstrates a significant improvement of 2.13% in IoU score and 0.96% in F1 score compared to the Relation distillation method, which are crucial evaluation metrics for assessing the performance.

According to Table 3, Table 5, and Table 7, even when applied to the most recent methods for building and road extraction, our distillation strategy continues to prove its effectiveness. Compared to the individually trained student network, IoU score has improved by 3.58%, 4.3% and 3.45% on the three datasets, respectively.

Table 7
Distillation performance of SOTA building and road extraction on the WHU Building Dataset.

Method	Precision	Recall	IOU	F1
CRAE-Net (T)	0.9582	0.9477	0.8878	0.9456
SGCN (T)	0.9568	0.9489	0.8791	0.9522
BiSeNetV2(S)	0.8582	0.8882	0.7726	0.8693
+Ours	0.9002	0.9098	0.8135	0.9028

5. Discussion

5.1. Visualization Analysis

In this section, we compare the extraction results of the proposed distillation strategy with the other three distillation strategies on the Massachusetts Road Dataset, LRSNY Road Dataset, and WHU Building Dataset to further compare and analyze the advantages and limitations of our proposed distillation strategy.

As shown by a few red boxes in Fig. 7, on the Massachusetts Roads Dataset, our approach captures and extracts road items with good accuracy. Compared to the extraction results of the original student model, our proposed distillation strategy achieves better completeness in road extraction, as demonstrated by the maximum display of effectiveness in the sixth row of extraction results. As shown by a few red boxes in Fig. 8, on the LRSNY Roads Dataset, our distillation strategy also demonstrates outstanding road object extraction capabilities. The second row of extraction results shows that our distillation strategy can effectively avoid the extraction of erroneous targets, as opposed to other distillation methods, as indicated by the comparison in the red small boxes. In the fourth row of extraction results comparisons, the KD distillation method also exhibits issues with extracting incorrect targets. As shown by a few red boxes in Fig. 9, on the WHU Building Dataset, our distillation strategy performs exceptionally well compared to the other three distillation methods. The outcomes of extraction's first and second rows of comparison indicate that our distillation strategy significantly improves the completeness and precision of target extraction for building objects.

5.2. Ablation Experiment

This section focuses on assessing the efficacy of our distillation strategy's tactics for extracting construction and road goals. In order to do this, we carried out ablation experiments and assessed them using three datasets of remote sensing images.

According to Tables 8, 9, and 10, we combined existing distillation strategies. In the first three rows of the table, we combined pairwise the response-based knowledge, feature-based knowledge, and relation-based knowledge distillation strategies. These combinatorial distillation strategies allowed us to see a significant improvement in the performance of the student model. Nevertheless, compared to the distillation strategy proposed in the fifth row, our distillation strategy effectively guided the student model to achieve improved segmentation performance. To further assess the efficacy of our proposed distillation strategy, which involves the mutual exchange of feature information between teacher networks, we conducted ablation experiments comparing it with a traditional knowledge distillation method that directly utilizes feature knowledge between teacher networks. The student network model's IoU and F1 scores climbed to 47.19% and 64.11%, respectively, on the Massachusetts Roads Dataset, respectively. After introducing the method of mutual exchange of feature knowledge among teacher networks, IoU and F1 ratings for the student network model increased to 47.33% and 64.25%, respectively. The IoU score significantly improved on the LRSNY Roads Dataset, with a 0.78% increase compared to the traditional method of directly utilizing feature knowledge between teacher networks.

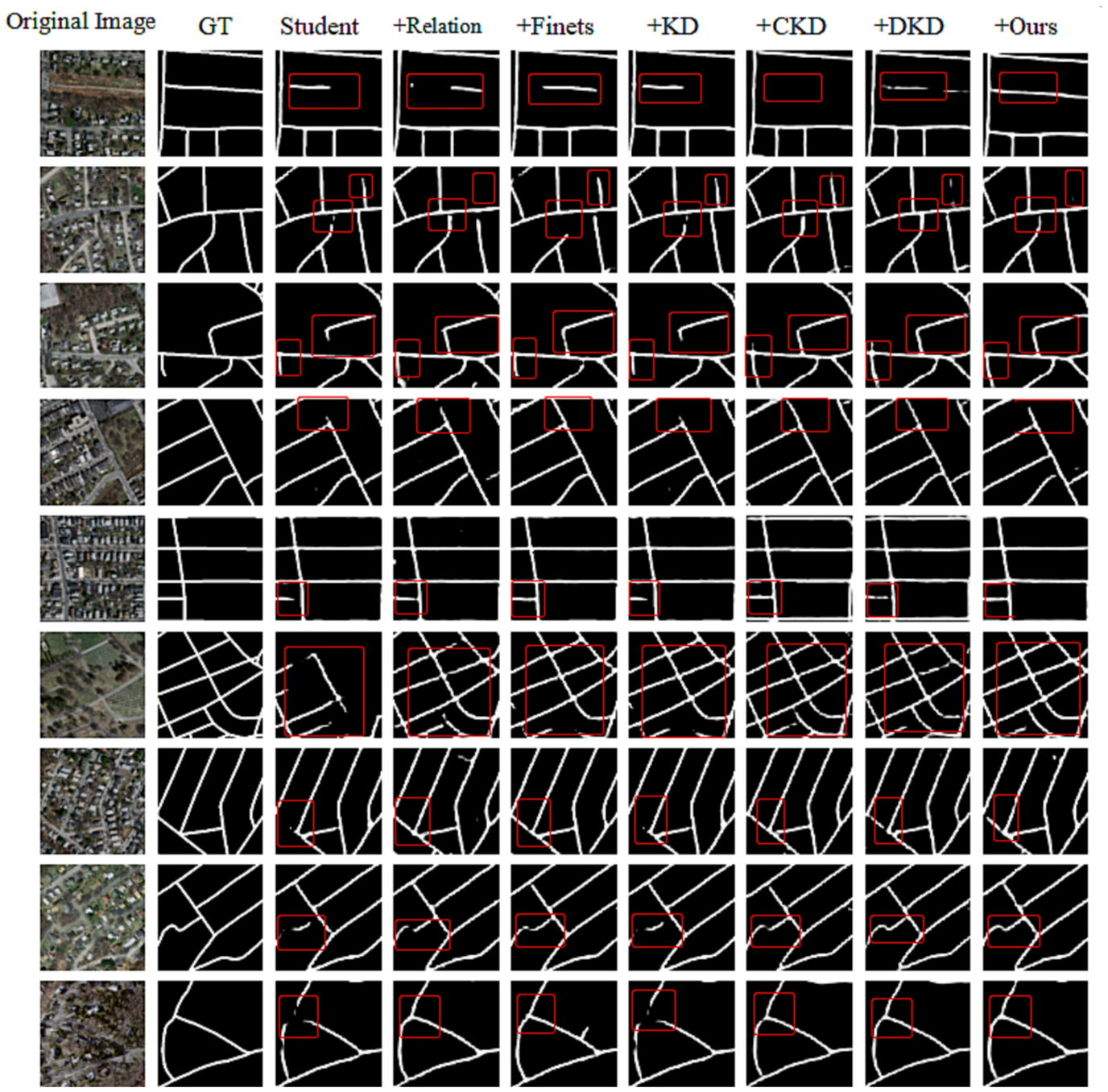


Fig. 7. Examples of extraction results on the Massachusetts Roads Dataset for our distillation strategy and several other distillation methods.

We particularly carried out ablation tests on the WHU Building Dataset to further confirm the efficacy of our suggested distillation technique. As depicted in Table 10. Our proposed distillation strategy exhibited superior performance compared to other combined distillation strategies, as evidenced by the highest IoU and F1 scores. These results serve as strong evidence of the effectiveness of our method. After introducing the method of mutual exchange of feature knowledge among teacher networks, the IoU score improved by 0.53%.

6. Conclusion

This paper introduces multi-teacher collaboration and multi-knowledge fusion mechanisms into knowledge distillation and applies them to road and building extraction tasks. The main innovations include multi-teacher collaboration distillation, multi-knowledge fusion, and the method of guiding the student network by exchanging feature knowledge among teacher networks. Through these innovations, the objective of this study is to enhance the performance of remote sensing image extraction tasks while reducing model size and computational resource requirements.

In this paper's approach, a strategy of multi-teacher collaboration distillation is employed. Different teacher models possess different strengths and expertise. Therefore, by combining multiple teacher models, their knowledge can be comprehensively utilized to provide more comprehensive guidance to the student model. Additionally, this paper introduces a mechanism of multi-knowledge fusion, which leverages knowledge from multiple teacher models to more fully utilize diverse knowledge and enhance the performance of the compact

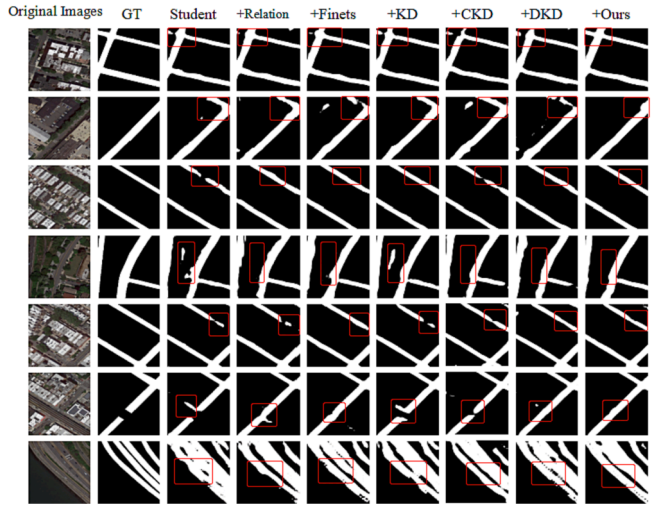


Fig. 8. Examples of extraction results on the LRSNY Roads Dataset for our distillation strategy and several other distillation methods.

network.

Finally, this paper proposes a method that guides the student network by exchanging feature knowledge among teacher networks. By exchanging feature knowledge among teacher networks, richer feature

Table 8
Ablation Experiment Analysis on Massachusetts Roads Dataset.

Method	Precision	Recall	IOU	F1
KD + Relation	0.6743	0.6012	0.4677	0.6403
KD + Finets	0.6823	0.6050	0.4711	0.6398
Finets + Relation	0.6832	0.5984	0.4697	0.6395
Ours(student feature)	0.6841	0.6031	0.4719	0.6411
Ours(teacher feature)	**0.6850**	**0.6062**	**0.4733**	**0.6425**

Table 9
Ablation Experiment Analysis on LRSNY Roads Dataset.

Method	Precision	Recall	IOU	F1
KD + Relation	0.8772	0.8662	0.7725	0.8716
KD + Finets	0.8758	0.8680	0.7729	0.8719
Finets + Relation	0.8760	0.8573	0.7642	0.8767
Ours(student feature)	0.8794	0.8655	0.7737	0.8725
Ours(teacher feature)	**0.8885**	**0.8685**	**0.7815**	**0.8774**

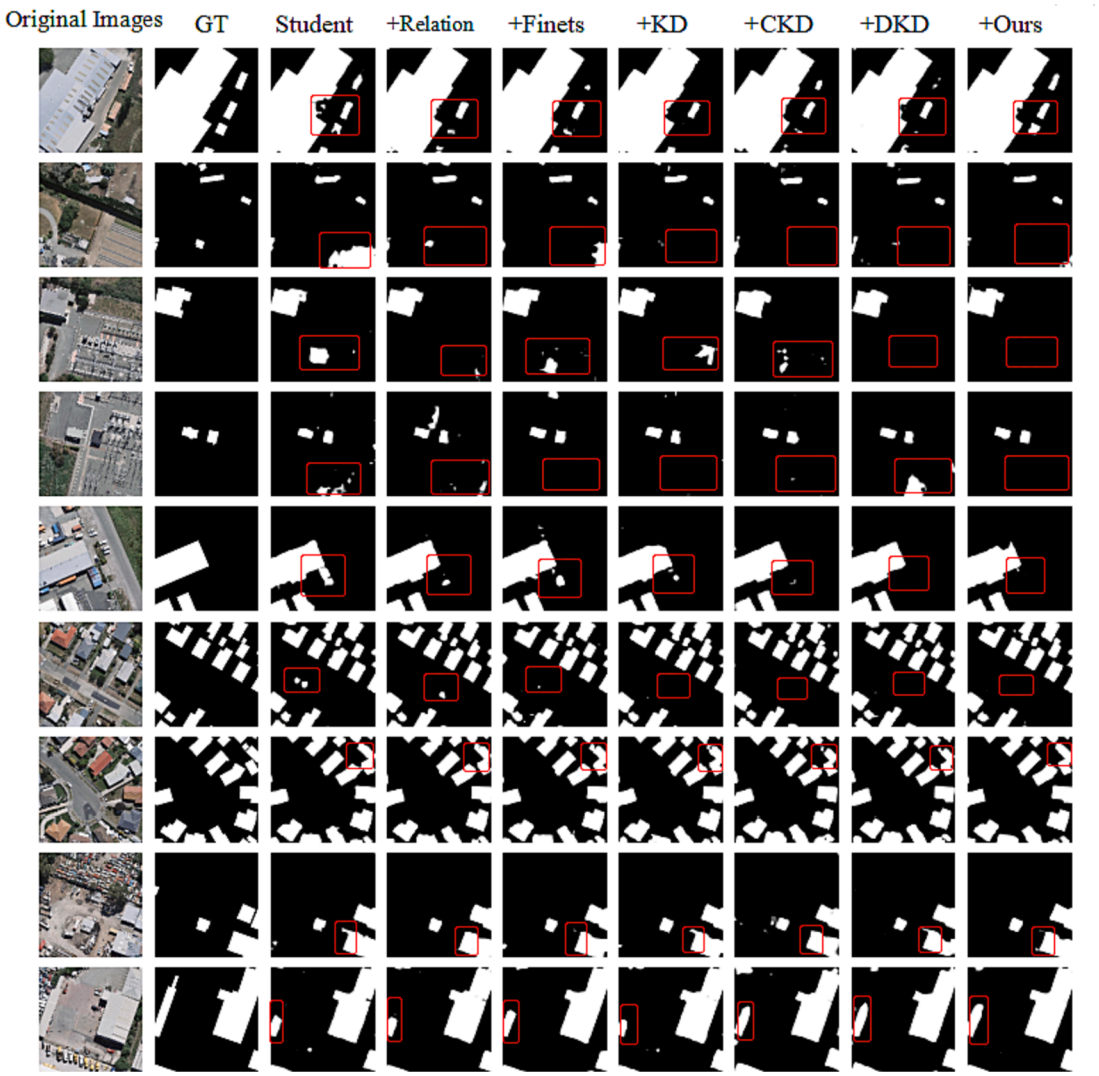


Fig. 9. Examples of extraction results on the WHU Building Dataset for our distillation strategy and several other distillation methods.

Table 10
Ablation Experiment Analysis on WHU Building Dataset.

Method	Precision	Recall	IOU	F1
KD + Relation	0.8821	0.8971	0.7950	0.8890
KD + Finets	0.8835	0.8966	0.8010	0.8930
Finets + Relation	0.8829	0.8989	0.7983	0.8917
Ours(student feature)	0.8833	0.8984	0.8018	0.8887
Ours(teacher feature)	**0.8847**	**0.9020**	**0.8071**	**0.8933**

knowledge can be transferred to the student model, guiding its training process. This feature-based knowledge adoption method improves the student model's understanding and extraction capabilities of image features.

Through rigorous practical evaluations of multiple state-of-the-art distillation procedures on various remote sensing image datasets, the recommended strategy in this study demonstrates significant performance gains. Our suggested distillation method outperforms other cutting-edge distillation techniques on the Massachusetts Roads Dataset with an amazing IoU increase of 0.39% and an F1 gain of 0.53%. On the LRSNY Roads Dataset, our distillation method outperformed other top-performing distillation strategies with a 1.14% IoU gain and a 0.85% precision gain. On the WHU Building Dataset, Our distillation technique surpasses other cutting-edge methods with a 1.19% IoU gain and a 0.34% precision gain.

CRediT authorship contribution statement

Ziyi Chen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Project administration, Funding acquisition. **Liai Deng:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft. **Jing Gou:** Resources, Writing – review & editing. **Cheng Wang:** Conceptualization, Supervision. **Jonathan Li:** Investigation, Formal analysis. **Dilong Li:** Formal analysis, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This study was financially supported by the Natural Science Foundation of Fujian Province (No. 2023J01135), National Natural Science Foundation of China (No.62001175), Fundamental Research Funds for the Central Universities of Huaqiao University (No.ZQN-911), the National Natural Science Foundation of China (No. 42201475, 61972168), the Natural Science Foundation of Fujian Province (NO. 2022J01317, 2021J05059), Fundamental Research Funds for the Central Universities of Huaqiao University (No.ZQN-1114), and in part by the Major Science and Technology Project of Xiamen (Industry and Information Technology Area) (NO.3502Z20231007).

References

Bagherinezhad, H., Horton, M., Rastegari, M., Farhadi, A., 2018. Label refinery: improving imagenet classification through label progression. ArXiv abs/1805.02641.

Bhat, P., Arani, E., Zonooz, B., 2021. Distill on the Go: Online knowledge distillation in self-supervised learning. In: Paper presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 19-25 June 2021. https://doi.org/10.1109/CVPRW53098.2021.00301.

Chen, L.-C., Zhu, Y., Papandreou, Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Paper presented at the European Conference on Computer Vision, Munich, Germany, September 8-14, 2018. https://doi.org/10.1007/978-3-030-01234-2_49.

Chen, D., Mei, J., Zhang, H., Wang, C., Feng, Y., Chen, C., 2022a. Knowledge Distillation with the Reused Teacher Classifier. In: Paper presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18-24 June 2022. https://doi.org/10.1109/CVPR52688.2022.01163.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018b. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRF. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.

Chen, C., Park, T., Wang, X., Piao, S., Xu, B., 2019. China and India lead in greening of the world through land-use management. Nat. Sustain. 122–129 https://doi.org/10.1038/s41893-019-0220-7.

Chen, H., Wu, B., Yu, B., Chen, Z., Wu, Q., Lian, T., Wang, C., Li, Q., Wu, J., 2021. A new method for building-level population estimation by integrating LiDAR, nighttime light, and POI data. J. Remote Sens. 1–17. https://doi.org/10.34133/2021/9803796.

Chen, H., Peng, S., Du, C., Li, J., Wu, S., 2022b. SW-GAN: road extraction from remote sensing imagery using semi-weakly supervised adversarial learning. Remote Sens. 14, 4145. https://doi.org/10.3390/rs14174145.

Chen, S., Shi, W., Zhou, M., Zhang, M., Xuan, Z., 2022d. CGSANet: a contour-guided and local structure-aware encoder–decoder network for accurate building extraction from very high-resolution remote sensing imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 1526–1542. https://doi.org/10.1109/jstars.2021.3139017.

Chen, J., Zhang, D., Wu, Y., Chen, Y., Yan, X., 2022c. A context feature enhancement network for building extraction from high-resolution remote sensing imagery. Remote Sens. 14 https://doi.org/10.3390/rs14092276.

Deng, J., Guo, J., Xue, N., Zafeiriou, S, 2018. ArcFace: Additive angular margin loss for deep face recognition. In: Paper presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15-20 June 2019. 10.1109/CVPR.2019.00482.

Furlanello, T., Lipton, Z.C., Tschannen, M., Itti, L., Anandkumar, A., 2018. Born again neural networks. Int. Conf. Mach. Learn.

Guo, H., Shi, Q., Du, B., Zhang, L., Wang, D., Ding, H., 2021. Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. 59, 4287–4306. https://doi.org/10.1109/tgrs.2020.3014312.

He, S., Jiang, W., 2021. Boundary-assisted learning for building extraction from optical remote sensing imagery. Remote Sens. 13, 760. https://doi.org/10.3390/rs13040760.

Heo, B., Lee, M., Yun, S., Choi, J.Y., 2018. Knowledge distillation with adversarial samples supporting decision boundary. In: AAAI Conf. Artif. Intell. abs/1805.05532. 10.1609/AAAI.V33I01.33013771.

Hinton, G.E., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. ArXiv abs/1503.02531.

Hosseinpour, H., Samadzadegan, F., Javan, F.D., 2022. CMGFNet: a deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. ISPRS J. Photogrammetry Remote Sens. 184, 96–115. https://doi.org/10.1016/j.isprsjprs.2021.12.007.

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 7, 14680–14707. https://doi.org/10.3390/rs71114680.

Ji, M., Seungjae, S., Seunghyun, H., Gibeom, P., Il-Chul, M., 2021. Refine Myself by Teaching Myself : Feature Refinement via Self-Knowledge Distillation. In: Paper presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20-25 June 2021. https://doi.org/10.1109/CVPR46437.2021.01052.

Kang, Z., Zhang, P., Zhang, X., Sun, J., Zheng, N., 2021. Instance-conditional knowledge distillation for object detection. Neural Inf. Process. Syst. https://doi.org/10.48550/arXiv.2110.12724.

Lei, Y., Yu, J., Chan, S., Wu, W., Liu, X., 2022. SNLRUX++ for building extraction from high-resolution remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 409–421. https://doi.org/10.1109/jstars.2021.3135705.

Li, Z., Ye, J., Song, M., Huang, Y., Pan, Z., 2021. Online Knowledge Distillation for Efficient Pose Estimation. In: Paper presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10-17 October 2021. https://doi.org/10.1109/ICCV48922.2021.01153.

Li, S., Liao, C., Ding, Y., Hu, H., Jia, Y., Chen, M., Xu, B., Ge, X., Liu, T., Wu, D., 2021a. Cascaded residual attention enhanced road extraction from remote sensing images. ISPRS Int. J. Geo-Information 11, 9. https://doi.org/10.3390/ijgi11010009.

Li, K., Wan, G., Cheng, G., Meng, L., Han, J., 2020. Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS J. Photogramm. Remote Sens. 159, 296–307.

Li, J., Zhuang, Y., Dong, S., Gao, P., Dong, H., Chen, H., Chen, L., Li, L., 2022. Hierarchical disentangling network for building extraction from very high resolution optical remote sensing imagery. Remote Sens. 14, 1767. https://doi.org/10.3390/rs14071767.

Lin, T.-Y., Dollar, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J., 2017a. Feature pyramid networks for object detection. In: Paper presented at the In: Paper presented

at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21-26 July 2017. https://doi.org/10.1109/CVPR.2017.106.
Lin, S., Xie, H., Wang, B., Yu, K., Chang, X., Liang, X., Wang, G., 2022. Knowledge Distillation via the Target-aware Transformer. In: Paper presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA , 18-24 June 2022. https://doi.org/10.1109/CVPR52688.2022.01064.
Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. 42, 318–327. https://doi.org/10.1109/TPAMI.2018.2858826.
Luo, L., Wang, J.-X., Chen, S.-B., Tang, J., Luo, B., 2022. BDTNet: road extraction by bi-direction transformer from remote sensing images. IEEE Geosci. Remote Sens. Lett. 19, 1–5. https://doi.org/10.1109/lgrs.2022.3183828.
Mnih, V., Hinton, G.E., 2010. Learning to detect roads in high-resolution aerial images. In: Paper presented at the European Conference on Computer Vision, Hersonissos, Greece, September 5-11, 2010. https://doi.org/10.1007/978-3-642-15567-3_16.
Park, W., Kim, D., Lu, Y., Cho, M., 2019. Relational knowledge distillation. In: Paper presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15-20 June 2019. https://doi.org/10.1109/CVPR.2019.00409.
Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y., 2014. FitNets: Hints for Thin Deep Nets. CoRR abs/1412.6550.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. ArXiv abs/1505.04597. https://doi.org/10.1007/978-3-319-24574-4_28.
Schrotter, G., Hürzeler, C., 2020. The digital twin of the city of Zurich for urban planning. J. Photogramm. Remote Sens. Geoinform. Sci. 88, 99–112. https://doi.org/10.1007/s41064-020-00092-2.
Shi, W., Miao, Z., Wang, Q., Zhang, H., 2014. Spectral-spatial classification and shape features for urban road centerline extraction. IEEE Geosci. Remote Sens. Lett. 11, 788–792. https://doi.org/10.1109/lgrs.2013.2279034.
Wang, L., Bai, X., Gong, C., Zhou, F., 2021a. Hybrid inference network for few-shot SAR automatic target recognition. IEEE Trans. Geosci. Remote Sens. 59, 9257–9269. https://doi.org/10.1109/TGRS.2021.3051024.
Wang, S., Mu, X., Yang, D., He, H., Zhao, P., 2021b. Road extraction from remote sensing images using the inner convolution integrated encoder-decoder network and directional conditional random fields. Remote Sens. 13, 465. https://doi.org/10.3390/rs13030465.
Wang, Y., Peng, Y., Li, W., Alexandropoulos, G.C., Yu, J., Ge, D., Xiang, W., 2022a. DDU-Net: Dual-Decoder-U-Net for road extraction using high-resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–12. https://doi.org/10.1109/TGRS.2022.3197546.
Wang, Y., Zeng, X., Liao, X., Zhuang, D., 2022b. B-FGC-Net: a building extraction network from high resolution remote sensing imagery. Remote Sens. 14, 269. https://doi.org/10.3390/rs14020269.
Wu, Q., Luo, F., Wu, P., Wang, B., Yang, H., Wu, Y., 2021. Automatic road extraction from high-resolution remote sensing images using a method based on densely connected spatial feature-enhanced pyramid. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 3–17. https://doi.org/10.1109/jstars.2020.3042816.
Xia, L., Zhang, X., Zhang, J., Yang, H., Chen, T., 2021. Building extraction from very-high-resolution remote sensing images using semi-supervised semantic edge detection. Remote Sens. 13 https://doi.org/10.3390/rs13112187.
Xu, G., Liu, Z., Li, X., Loy, C.C., 2020. Knowledge distillation meets self-supervision. In: Paper presented at the European Conference on Computer Vision, August 23-28, 2020. https://doi.org/10.1007/978-3-030-58545-7_34.
Yan, C.C., Biao, G., Yuxuan, W., Yue, G., 2021. Deep multi-view enhancement hashing for image retrieval. IEEE Trans. Pattern Anal. Mach. Intel. 43, 1445–1451. https://doi.org/10.1109/TPAMI.2020.2975798.
Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., Zhang, Q., 2022. Cross-Image Relational Knowledge Distillation for Semantic Segmentation. In: Paper presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18-24 June 2022. https://doi.org/10.1109/CVPR52688.2022.01200.
Yim, J., Joo, D., Bae, J.-H., Kim, J., 2017. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In: Paper presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21-26 July 2017. https://doi.org/10.1109/CVPR.2017.754.
You, Z.-H., Wang, J.-X., Chen, S.-B., Tang, J., Luo, B., 2022. FMWDCT: foreground mixup into weighted dual-network cross training for semisupervised remote sensing road extraction. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 5570–5579. https://doi.org/10.1109/jstars.2022.3188025.
Yuan, L., Tay, F.E.H., Li, G., Wang, T., Feng, J., 2020. Revisiting Knowledge Distillation via Label Smoothing Regularization. In: Paper presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13-19 June 2020. https://doi.org/10.1109/CVPR42600.2020.00396.
Zagoruyko, S., Komodakis, N., 2016. Paying more attention to attention_Improving the performance of convolutional neural networks via attention transfer. ArXiv abs/1612.03928.
Zhang, H., Chen, D., Wang, C., 2021. Confidence-Aware Multi-Teacher Knowledge Distillation. In: Paper presented at the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23-27 May 2022. https://doi.org/10.1109/icassp43922.2022.9747534.
Zhang, L., Chen, X., Tu, X., Wan, P., Xu, N., Ma, K., 2022. Wavelet Knowledge Distillation: Towards Efficient Image-to-Image Translation. In: Paper presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18-24 June 2022. https://doi.org/10.1109/CVPR52688.2022.01214.
Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J., 2022. Decoupled Knowledge Distillation. In: Paper presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18-24 June 2022. https://doi.org/10.1109/CVPR52688.2022.01165.
Zhou, G., Chen, W., Gui, Q., Li, X., Wang, L., 2022a. Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–15. https://doi.org/10.1109/tgrs.2021.3128033.
Zhou, Y., Chen, Z., Wang, B., Li, S., Liu, H., Xu, D., Ma, C., 2022c. BOMSC-Net: boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 60, 1–17. https://doi.org/10.1109/tgrs.2022.3152575.
Zhou, M., Sui, H., Chen, S., Liu, J., Shi, W., Chen, X., 2022b. Large-scale road extraction from high-resolution remote sensing images based on a weakly-supervised structural and orientational consistency constraint network. ISPRS J. Photogrammetry Remote Sens. 193, 234–251. https://doi.org/10.1016/j.isprsjprs.2022.09.005.