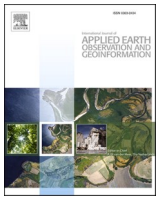
International Journal of Applied Earth Observations and Geoinformation 107 (2022) 102677

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



Road marking extraction in UAV imagery using attentive capsule feature pyramid network

Haiyan Guan [a,*], Xiangda Lei [a], Yongtao Yu [b], Haohao Zhao [a], Daifeng Peng [a], José Marcato Junior [c], Jonathan Li [d]

[a] School of Remote Sensing & Geomatics Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China
[b] Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223003, China
[c] Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Campo Grande 79070900, Brazil
[d] Department of Geography and Environmental Management and Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Keywords:
Road markings
Capsule
Feature pyramid network
Dense atrous convolution
UAV images

ABSTRACT

Accurately and precisely delineating road-markings from very high spatial resolution unmanned aerial vehicle (UAV) images face many challenges, such as complex scenarios, diverse road marking sizes and shapes, and absent and occluded road markings. To address these issues, we formulate an attentive capsule feature pyramid network (ACapsFPN) by integrating capsule representations with attention mechanisms into the feature pyramid network (FPN), aiming at improving road marking extraction accuracy. Different from the current convolutional neural network (CNN) models based on scalar neuron representations, capsule networks characterize entity features by leveraging vectorial capsule neurons, whose lengths and instantiation parameters contribute to the identification of features and their variants. By constructing a capsule FPN, the ACapsFPN is capable of extracting and integrating multi-level and multi-scale capsule features to provide high-quality and semantically-strong feature abstractions. By formulating a multi-scale context feature descriptor and the ternary feature attention modules, the ACapsFPN can emphasize informative features to generate a class-specific feature representation. Quantitative and qualitative evaluations show the ACapsFPN provides a valuable means for extracting road markings in UAV images under different kinds of complex conditions. In addition, comparative analyses with existing alternatives also demonstrate the superiority and robustness of the ACapsFPN in UAV road marking extraction.

1. Introduction

Every day, road markings are used as an efficient and indispensable means to provide millions of road users, i.e., pedestrian and car drivers, with guidance and protection. Automated detecting road markings has become an increasing necessity for transportation-related activities, including traffic monitoring, automatic vehicle driving, and autonomous navigation (Tian et al., 2018; Zhang et al., 2018). The fundamental objective of road marking detection is to provide shape and location information of individual road markings at centimeter-level accuracy for lane-based models and high-definition (HD) maps (Azimi et al., 2019). At present, most HD maps are generated for autonomous driving by LiDAR, Radar, global positioning system (GPS), or image/vision sensors mounted on land-based mobile mapping platforms (McCall and Trivedi, 2006; de Paula and Jung, 2015; Gupta and Choudhary, 2018; Wen et al., 2019; Xu et al., 2021).These methods come with the following drawbacks: (1) road marking data missing occluded by traffic flow and limited by the sensor line of sight, and (2) decreased mapping accuracy caused by global positioning system (GPS) signal loss in urban canyons (Azimi et al., 2019).

The compact and light-weighted unmanned aerial vehicle (UAV) is a trend for future earth observation data acquisition due to its cost saving, high efficiency, operational convenience for image retrieval (Lyu et al., 2020). Compared to satellite and aerial images, UAV images have very-high spatial resolutions, e.g., a 50-cm ground sampling distance (GSD), which provides more promising opportunities for cadastral mapping, and agriculture-related applications (e.g., smart farming, precision agriculture, and weed monitoring). Recently, the UAV technology has been used in a variety of applications in the transportation-related fields, ranging from traffic network monitoring, population density

* Corresponding author.
E-mail address: guanhy.nj@nuist.edu.cn (H. Guan).

https://doi.org/10.1016/j.jag.2022.102677
Received 28 October 2021; Received in revised form 23 December 2021; Accepted 6 January 2022
Available online 22 January 2022

monitoring, infrastructure monitoring, urban greenery monitoring to road safety improvement. Therefore, UAV images, for the purpose of HD map creation, are promising for extracting and classifying road markings.

At present, road marking detection or extraction from images is still a challenging problem (Fig. 1). The difficulties are summarized as follows.

(1) Road markings appear discontinuities and variations in shape, intensity, and size. Specifically, because road markings, as one of the essential road elements in transportation management systems, are critical for guiding pedestrians and drivers, they can be grouped into several categories, e.g., forbidden, instructional, and indicative, each of which contains different classes, e.g., single and double boundary, dotted and solid line, zigzag, pedestrian crossing, circular reflector, and speed limit. In terms of road marking size, road markings strongly rely on the images with different spatial resolutions. Moreover, many road markings are damaged after years of use, which leads to being partially and completely disappeared on the images due to decreased road marking reflectivity.
(2) Complex and highly-variable road scenarios. Due to the presence of objects e.g., bridges, trees, and a variety of vehicles, some road markings are partially or fully occluded. Lightning condition has important functions in the quality and consistency of road markings on the images. Moreover, shadows coming from high-rise buildings and trees cause different illumination over road markings, and further change their spectral changes on the images.
(3) Sensor (i.e., camera) parameters are varying with different imaging systems (Ye et al., 2020). Some certain perspective distortions of the sensed images might be caused by different imaging sensor types, different mounting positions of the sensors, and different flying heights. Moreover, imaging sensor parameters are more or less varying during data acquisition, which further causes spectral inconsistencies among the sensed images.

To address this challenging task, an increasing number of road marking extraction methods have been proposed. Most traditional methods (e.g., Random Sample Consensus (RANSAC), Hough Transform (HT), and clustering) segment road markings from the images according to hand-crafted low-level features, e.g., spectral, textural, and geometrical features (Son et al., 2015; Jung et al., 2016; Niu et al., 2016; Li et al., 2018). Although many achievements have been obtained with improved accuracies and less computational complexity in specific situations, accurately and robustly extracting road markings in many complex road scenes is still challengeable, which cannot meet the requirements of HD map creation for autonomous driving (Xiao et al., 2020).

A variety of deep learning networks (e.g. convolutional neural network (CNN)) have drawn the increasing attention of researchers to effectively and highly-accurately detect, extract, and classify road networks and objects above or on road surfaces in complex road scenes, due to their powerful high-order feature representation, characterization, and robustness abilities (Xiao et al., 2020). However, most CNNs usually fail to extract heterogeneous object regions, and thus generating rough segmentation boundaries. Moreover, CNNs suffer from the issues of representation power and computational efficiency.

We propose a new attentive capsule feature pyramid network (ACapsFPN) that accurately and precisely extracts road markings from UAV images. The proposed ACapsFPN characterizes high-order entity features by leveraging vectorial capsule neurons. The ACapsFPN architecture includes: (1) a hierarchical encoder, which extracts multi-level information-rich capsule features at different scales, (2) a decoder integrating with lateral connections, which aggregates multi-scale capsule features to accurately extract road markings in UAV images, and (3) to strengthen the abstraction capability of the output features, a multi-scale context feature descriptor and ternary feature attention

Fig. 1. Challenges in road marking segmentation. For example, shadows caused by buildings and trees; partial occlusion caused by other objects, such as vehicles; severely destroyed road markings; false road markings caused by illumination, imaging tilt angles, and other road marking materials; presence of several types of road markings with varied sizes and shapes (e.g., turn signs, pedestrian crossing, zigzag, bus and bike sign, solid and dot line lanes).

modules are formulated and embedded into the ACapsFPN. The ACapsFPN provides a promising and competitive extraction performance of road markings with different spatial distributions and geometric topologies, varying intensity appearances and sizes, and diverse shapes and environmental scenarios in UAV images. Our contributions are listed as follows.

- We construct a capsule feature pyramid network (CapsFPN), which extracts and integrates multi-level and multi-scale capsule features to provide a high-quality task-aware feature semantics for improving road marking extraction performance.
- We design a multi-scale context feature (MCF) descriptor and ternary attention modules, i.e., a feature channel attention (FCA) module, a class region attention (CRA) module, and a class channel attention (CCA) module. Specifically, the MCF module aims to obtain multi-scale contextual information with no loss of feature details and resolutions. FCA aims to enhance the capability of feature representation by increasing the sensitivity of the network to information-rich and salient features. The CRA and CCA modules consider the spatial features and the importance of the feature channels, respectively, tightly related to the road markings, suppressing effectively the influences of the background and offering highly accurate road marking feature representations.

2. Related work

Recently, the detection and extraction of road markings has been attracted increasing attention in Intelligent Transportation Systems (ITS). We briefly review the existing image-based road marking extraction works. In this section, in terms of features to be used, current road marking studies have been roughly classified into two categories: traditional and deep learning-based methods.

2.1. Traditional road marking detection

Some systems were designed for detecting specific road markings, rather than all types of road markings (McCall and Trivedi, 2006). For example, to maintain vehicles to run along the host lanes, lane markings were considered for automatic vehicle driving and advanced driver assistant system (Huang et al., 2017). Because lane markings are characterized by linear features, many traditional image processing methods have been applied to lane marking detection, such as classical edge detection (e.g., Sobel and Canny detectors), template matching, Hough Transform, and threshold segmentation methods (e.g., local adaptive threshold segmentation method and Otsu's method). These methods mainly used intensity, texture, edge, geometric shape, and other low-level features to detect lane markings from images (de Paula and Jung, 2015; Lee and Moon, 2018). To deal with the variations of lane markings, more hand-crafted features have been explored, including Haar-like (Han et al., 2009), local binary pattern (Grabner et al., 2008), and dense vanishing point estimation (Ozgunalp et al., 2017). Moreover, by means of other data sources, such as depth information and the OpenStreetMap (OSM), some interference objects, e.g., vehicles and buildings, have been removed from the road scene to be processed, which improves the extraction of lane markings by coupling with other imaging processing algorithms (Prakash et al., 2015). To achieve highly-accurate lane detection results, some researchers proposed hierarchical lane detection methods (de Paula and Jung, 2015). Machine learning-based methods, e.g. support vector machine (SVM) (Kim, 2008) and random forest (RF) (Gopalan et al., 2012), have been employed for improving the detection accuracy of lane markings.

To best read the road for an autonomous vehicle, besides lane markings, more marking types, such as zebra crossings, intersections, arrows, painted traffic signs, should be read, classified, and interpreted (Mathibela et al. 2015). Mathibela et al. (2015) proposed a road marking classification framework by integrating geometric feature functions with probabilistic RUSBoost and Conditional Random Field (CRF), which jointly classified seven classes of road markings (i.e., single boundary, double boundary, zig-zag, separator, intersection, boxed junction, and special lane) with a precision of between 74% and 93% and a recall of between 69% and 94% across all classes. Some researchers hierarchically detected and classified road markings based on the characteristics of road markings. For example, symbol-based road markings were recognized by embedding Histogram of Oriented Gradient (HOG) features into the SVM, while text-based markings were recognized using Optical Character Recognition (OCR) (Greenhalgh and Mirmehdi, 2015). Gupta and Choudhary (2018) first clustered the image into lane marking clusters and non-lane marking clusters via a spatio-temporal incremental clustering (STIC) algorithm coupled with curve-fitting, and then classified non-lane marking clusters into road markings by a Grassmann manifold learning framework. However, these methods were much suitable for traffic scenes with good illumination conditions (Huang et al. 2017).

2.2. Deep learning-based road marking detection

Comparatively, deep learning-based road marking detection methods usually achieved better detection accuracies because these methods usually learn semantic, high-order features rather than using low-level, simple, and hand-crafted features (Huang et al. 2017; Azimi et al. 2019).

Hoang et al. (2019) proposed a road marking detection and classification framework by combining a simple, and hand-crafted feature-based method with a deep learning-based method. Specifically, the framework first created the region of interest (ROI) images via a vanishing point strategy, and further detected arrows and bike markings by the CNN-based detector and classifier. Lee et al. (2017) detected some road marking types under adverse weather conditions via a vanishing point guided net (VPGNet). Li et al. (2017) detected road lane boundaries by a jointed strategy, in which the multi-task CNN model provided the geometric information of road lanes, and the recurrent neural network automatically detected the boundaries of the road lanes. Moreover, a dual-view CNN was performed for detecting road lanes on the raw front-view image and its converted top-view image (He et al., 2016). The Region Convolutional Neural Network (R-CNN) and its variant, the faster R-CNN, were developed to detect small objects (Girshick et al., 2014; Girshick, 2015). Based on the faster R-CNN, Tian et al. (2018) detected lane markings by combining with fast multilayer feature maps, context information, and an anchor generating method. However, the method failed to process over-exposed images and recognize road markings occluded by the other objects, e.g., pedestrians, vehicles, trees, and buildings, on the road. Although deep learning models have made great achievements, they still suffer from the issues of voluminous training data required for model construction. That is, the quantity and the quality of the training data directly determine the effectiveness and robustness of the constructed models.

Generally speaking, traditional CNNs, constructed based on scalar neuron representations, represent the probabilities of the presence of specific features. To effectively capture the variances of an entity, a traditional CNN requires more extra neurons to respectively encode the different variants of the entity with the same type, which results in the expansion of the network size and parameters. Recently, capsule networks have shown superior performances on the capabilities of feature abstraction and representation. Unlike traditional CNNs, capsule networks characterize entity features by leveraging vectorial capsule neurons. Specifically, for a capsule neuron, its length encodes the probability of the existence of an entity, and its instantiation parameters describe the inherent properties of the entity (Sabour et al., 2017). Such a capsule formulation allows a capsule to not only detect a feature, but also to learn and identify its variants, resulting in a powerful but light-weight feature abstraction model. Capsule networks have shown promising performance in a set of prediction, detection, segmentation,

and classification tasks (Paoletti et al., 2019; Yu et al., 2021; Ma et al., 2021a,b). However, capsule networks have not been applied to road marking detection.

Feature pyramid, a top-down multi-scale feature fusion structure (Lin et al., 2017), offers unique features for object objection and segmentation. It combines high-level coarse semantic features and low-level location information to generate stronger feature representation (Shamsolmoali et al., 2021a,b,c). Subsequently, several Feature pyramid architectures have been proposed to improve object-detection performances, such as a cascaded pyramid network formulation (Chen et al., 2018), a multipatch feature pyramid network for weakly supervised object detection (MPFP-Net) (Shamsolmoali et al., 2021a), and a rotation equivariant feature image pyramid network (REFIPN) for efficiently detecting objects with size variations (Shamsolmoali et al., 2021b). Thus, feature pyramid has been increasingly used and modified for extracting variedly- sized and shaped specific objects, such as roads (Shamsolmoali et al., 2021a,b,c), buildings (Zhu et al., 2021), waterbodies (Yu et al., 2021), and road markings (Chen et al., 2021). To solve the specific problems raised by road markings of various sizes, we adopt a strategy similar to the FPN, where we use vectorial capsule convolution networks rather than traditional scalar convolution networks. Moreover, Attention mechanism has been attracted much attention because it enlarges the receptive field size. Channel attention emphasize important feature maps by calculating the weight for each channel (Hu et al., 2018). Xiao et al. (2020) detected lane markings by introducing self-attention and channel attention to capture global contextual information and strengthen important features. Moreover, the attention mechanism effectively improved the network performance without heavy computational costs (Hu et al., 2018). Thus, we propose the ACapsFPN, which fuses the attention modules into the capsule FPN for improving the extraction robustness of road markings.

3. ACapsFPN framework

3.1. Capsule network

Denote $\boldsymbol{u}_i$ as the vector output of a capsule i in the lower capsule convolutional layer, $\boldsymbol{v}_j$ as the vector output of capsule j in the capsule layer above. The implementation of the capsule network is detailed as follows.

(1) A prediction vector $\boldsymbol{U}_{j|i}$ is produced by:

$$U_{j|i} = \boldsymbol{W}_{ij} \cdot \boldsymbol{u}_i \tag{1}$$

where $\mathbf{W}_{ij}$ is a transformation matrix on the edge connecting capsules i and j. $\boldsymbol{U}_{j|i}$ is the prediction from capsule i to capsule j.

(2) A weighted sum, $\boldsymbol{s}_j$, the total input to capsule j, over all prediction vectors $\boldsymbol{U}_{j|i}$ is calculated from the capsules in the lower layer by:

$$s_j = \sum_i a_{ij} \cdot \boldsymbol{U}_{ji} \tag{2}$$

where, a_{ij} is the coupling coefficient, which indicates the degree of contribution of the prediction from capsule i to capsule j in the layer below. Note that the coupling coefficients a_{ij} sum to 1. Specifically, in each capsule layer, a_{ij} is designed as learnable parameters during network training, rather than determined by the dynamic routing process (Kim, 2008), because they are unstable and difficult to converge for the deep networks.

(3) A nonlinear "squashing" function (Kim, 2008) acts as the activation function to regulate the output of capsule $\boldsymbol{s}_j$. The squashing function is defined by:

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \cdot \frac{s_j}{\|s_j\|} \tag{3}$$

Through modulation, short capsules are suppressed to almost a length of almost zero to cast low predictions, whereas long capsules are compressed to a length of nearly one to cast high predictions.

3.2. Attentive capsule feature pyramid network

3.2.1. Overview

With the advantages of capsule networks and attention mechanisms, we formulate an attentive capsule feature pyramid network (named ACapsFPN) to obtain better road marking extraction performance. As illustrated in Fig. 2, the ACapsFPN, which is designed as a fully convolutional capsule feature pyramid network architecture, inputs a UAV image and outputs a road marking map with the identical image size in an end-to-end manner. The ACapsFPN is composed of a hierarchical encoder, a decoder, and several lateral connections. Specifically, the encoder aims to extract information-significant capsule features at multiple levels and scales. The decoder and the lateral connections take charge of aggregating the capsule features to generate a high-quality task-aware feature encoding, and finally obtain a highly accurate road marking map. Additionally, to further enhance the feature abstraction capability and class-aware feature encodings, we construct a multi-scale context feature descriptor and three types of feature attention modules, and embed them into the capsule feature pyramid network in place.

(1) **Hierarchical encoder**

In the hierarchical encoder (see Fig. 2), two scalar traditional convolutional layers with 256 convolution kernels are first used to extract low-level road marking features from the input image (Denote H and W as the height and width of an input image). Note that the rectified linear unit (ReLU) is adopted for the two traditional convolutional layers. Afterwards, the primary capsule layer is constructed to encode the low-order scalar feature outputs as high-order vectorial capsule neurons. All traditional and capsule convolutional layers are used with the convolution kernel size of 3×3, stride of 1, and padding of 1. Denote D_p and S_p as the number of feature channels and the dimension of a capsule, respectively. In the primary capsule layer, a total of $D_p \times S_p$ kernels are designed to slide on the second traditional convolutional layer, generating $D_p \times S_p$ capsule feature channels. In other words, the generated capsule feature channels are equally partitioned into D_p groups, each of which includes S_p feature channels. In such way, for each group, the S_p elements at the same position across the feature channels are concatenated to constitute an S_p-dimensional capsule representation. In this study, we set $D_p = 64$ and $S_p = 16$, respectively.

That is, each of 64 groups in the primary capsule layer finally forms a capsule feature channel with 16 dimensions.

The hierarchical encoder is subsequently designed with four network stages, each of which contains seven capsule convolutional layers for generating feature maps and a capsule max-pooling layer (except the last stage) for gradually scaling down with a scale of 2. Similarly, for all capsule layers in the four stages, we set $D_p = 64$ and $S_p = 16$, respectively. Specifically, in the first stage, the output of the primary capsule layer is put through seven 3×3 capsule convolution layers, followed by a 2×2 capsule max-pooling layer with a stride of 2. The output feature maps from the seven capsule convolution layers have the same spatial size with the input (i.e. $H \times W$). The output from the max-pooling layer is scaled down to the half size (i.e. $H/2 \times W/2$), while highlighting the most representative features. Stage by stage, the spatial resolution of the feature maps decrease gradually, whereas their feature abstraction level is higher.

Accordingly, the output from the deepest capsule layer in each stage has the highest-order feature representation, and is selected to build a

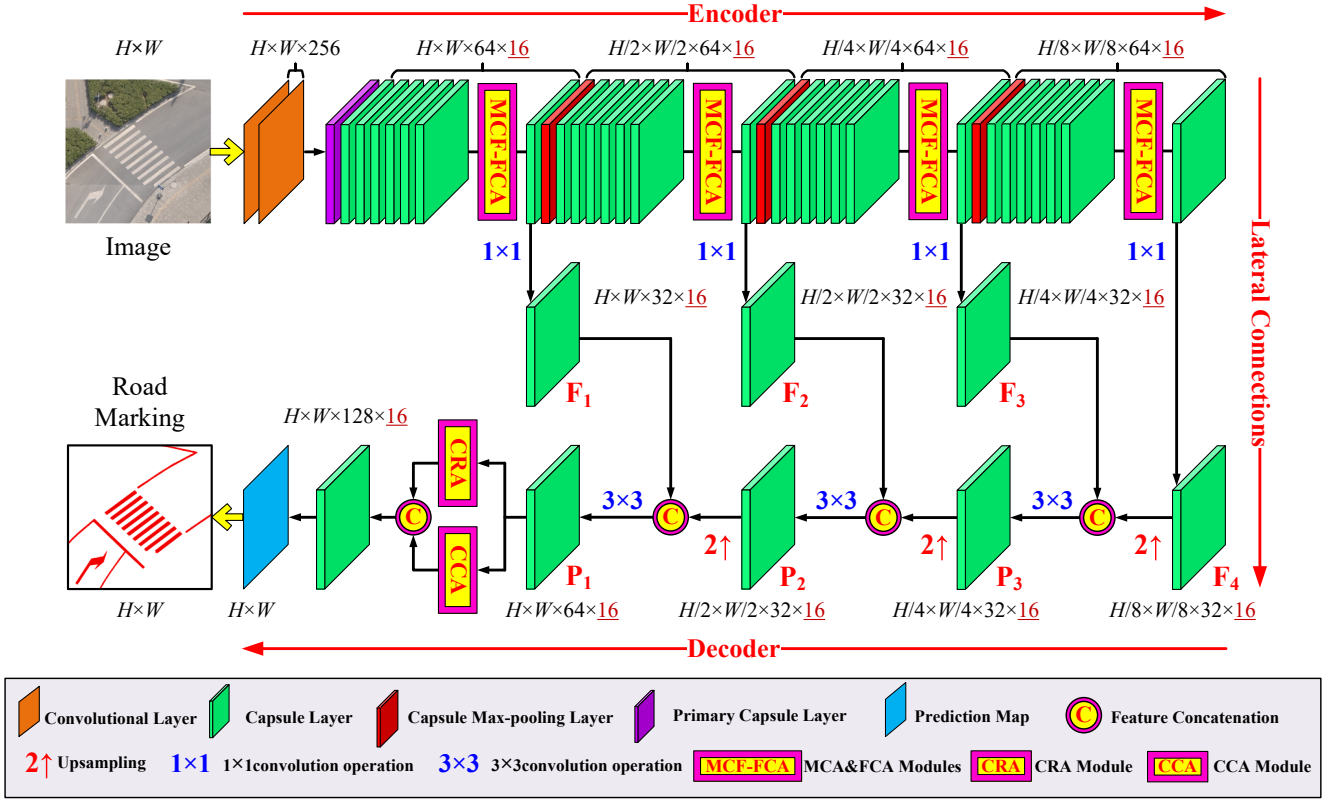


Fig. 2. Architecture of the proposed ACapsFPN.

feature map reference set. To effectively fuse features and reduce the number of network parameters, a 1×1 capsule convolution layer is applied to each feature map in the reference set to modulate their channel numbers to the same corresponding configuration while maintaining their original spatial resolutions. In other words, after the 1×1 capsule convolution operation, the number of feature channels, $D_p = 64$, is reduced to half previous size, $D_p = 32$. After feature map selection and modulation, the feature maps (i.e., $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$, and $\mathbf{F}_4$) with the corresponding scales of 1, 1/2, 1/4, and 1/8, respectively, regarding the input image, are finally selected as the feature map reference set for the subsequent feature fusion and enhancement.

However, max-pooling operations partially damage the details in the feature maps with the lower spatial resolution. Moreover, because convolutional operation is equally applied to all the channels of a capsule feature map at each layer, leading to insufficient use of the saliencies among the channels to obtain information-rich features and suppressing the channels less useful for road marking prediction. To address this issue, we integrate a multi-scale context feature (MCF) descriptor for feature augmentation and a feature channel attention (FCA) module for feature recalibration, respectively, over the deepest capsule layer in each stage (see Fig. 2). By embedding the MCF descriptor (see Section 3.2.2) into each stage, the output can effectively include contextual information at different scales without reducing road marking feature resolutions and details. With the FCA module (see Section 3.2.3) embedded into each stage, the informative road marking features are effectively highlighted, whereas the less salient features are suppressed, thereby further strengthening the road marking feature representation capability of the capsule feature maps.

(2) **Decoder & lateral connections**

To improve road marking extraction accuracy, the decoder is cooperatively worked with the lateral connections to integrate the multi-level and multi-scale features in $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$ to generate a high-quality task-oriented feature semantics. As shown in Fig. 2, first, to facilitate feature concatenation and feature fusion, the feature map along the decoder (i.e., along with the red arrow) is spatially scaled up with a scale of 2, which is achieved by capsule deconvolutional operations with a kernel size of 3×3, a stride of 1, and a padding of 1. For example, in the feature map reference set, feature map F4 (with the scale of 1/8) is upsampled with a scale of 2, and concatenated with feature map F3 (with the scale of 1/4) through the lateral connection.

Then, the concatenated feature map is leveraged to perform feature fusion by 3×3 capsule convolutional operation, which results in a high-resolution and semantically-strong feature representation, feature map P3. The above feature fusion process repeats downward to gradually fuse all the reference feature maps in $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$, generating three feature maps $\{\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3\}$. Note that, for the first two feature fusion processes, feature maps, $\mathbf{P}_1$ and $\mathbf{P}_2$, maintain their channel numbers to the same corresponding configuration of the feature map reference set, that is $D_p = 32$. In the last feature fusion process, feature map P1, which combines all the scales of features, is finally processed to predict the road marking map with $D_p = 64$. In fact, feature map P1 includes a global feature encoding for the whole input image. The spatial features of the road markings are not explicitly highlighted and the background features are not rationally reduced. Additionally, the feature channels of $\mathbf{P}_1$ that are tightly related to road markings are not positively emphasized, either. Therefore, it is not powerful enough to directly apply feature map P1 to obtain a high-quality road marking prediction map. Thus, to improve road marking extraction accuracy, we design two types of class-specific attention modules over the feature map P1, i.e., a class region attention (CRA) module and a class channel attention (CCA) module (see Section 3.2.3), respectively, to pay close attention to the spatial features of road markings and highlight the channels of the road marking features. As shown in Fig. 2, the outputs by the CRA module and the CCA module, respectively, are concatenated to form a powerful feature representation for generating a road marking prediction map.

3.2.2. Multi-scale contextual feature (MCF) descriptor

The MCF descriptor encodes the high-level semantic feature maps by leveraging atrous convolution operations [45]. An atrous rate indicates the stride of the convolution kernels in atrous convolutions. By adjusting the atrous rate, atrous convolutions can access different-size receptive fields with no increase of the number of kernel parameters. Assuming that the input capsule feature map,$\boldsymbol{I}_{\mathrm{MCF}} \in \boldsymbol{R}^{H\times W\times 64\times 16}$, contains 64 16-dimensional capsule features with the size of $H \times W$ pixels. As shown in Fig. 3, the MCF descriptor is designed as a four-parallel-branch structure, by which an augmented feature map (i.e., H × W × 64 × 16) is generated to explore multi-scale contextual information with the designed atrous rates. Concretely, the atrous rates of 1, 2, and 3 are given for the first braches, respectively, and the atrous rates of 3 and 5 for the last branch. With the given atrous rates, the four braches encapsulates small, middle, and large ranges of contextual information, respectively, by 3 × 3 atrous convolution operations. Afterwards, a 1 × 1 atrous convolution with an atrous rate of 1 is performed on each branch for modulating the convoluted feature channels. Along with the original input feature, the four-parallel-branch outputs are concatenated and fused through a 1 × 1 atrous convolution to finally output the feature map, $\boldsymbol{O}_{MCF} \in \boldsymbol{R}^{H\times W\times 64\times 16}$, which encapsulates multi-scale contextual information without reducing feature details and resolutions.

3.2.3. Ternary attention modules

(1) **Feature channel attention (FCA) module**

The FCA module is designed to model the interdependencies among the channels for strengthening feature saliency and weakening the features unconducive to prediction with a global perspective. As shown in Fig. 4, a road marking feature map input, $\boldsymbol{I}_{\mathrm{FCA}} \in \boldsymbol{R}^{H\times W\times 64\times 16}$, also contains 64 16-dimensional capsule features with the size of $H \times W$ pixels. Concretely,

(1) A 1 × 1 capsule convolution operation is applied to the road marking feature input, $\boldsymbol{I}_{\mathrm{FCA}}$, to generate a one-dimensional capsule feature map, $\boldsymbol{A} \in \boldsymbol{R}^{H\times W\times 64}$, containing 64 channels with the identical spatial size of the road marking feature map input, $\boldsymbol{I}_{\mathrm{FCA}}$. By the 1 × 1 convolution operation, the probability properties of $\boldsymbol{I}_{\mathrm{FCA}}$ are encoded to further model the interdependencies of the channels.
(2) A global average-pooling operation is performed on feature map $\boldsymbol{A}$ to generate a channel descriptor, each channel of which can be defined by:

$$a_i = \frac{1}{H\times W}\sum_j \left\| U_j^i \right\| \tag{4}$$

where $a_i$denotes the squeezed value associated with the i-th channel of $\boldsymbol{A}$; U_j^i is the j-th capsule in the i-th channel of $\boldsymbol{I}_{\mathrm{FCA}}$. Through the global average-pooling operation, a scalar value is computed by spatially squeezing the lengths of the capsules. The obtained channel descriptor contains 64 channels, identical to the number of channels of feature map $\boldsymbol{A}$, each of which correspondingly characterizes a global perspective of the feature statistics of that channel.

(3) Afterwards, two fully-connected (FC) convolution operation are applied to the channel descriptor to exploit the interdependencies within channels in a non-mutually exclusive manner, followed by the two activation functions of the ReLU and the sigmoid, respectively. Thus, the output obtained from the second fully-connected operation encodes the importance probabilities of the channels to form a channel-wise attention descriptor, denoted as $\boldsymbol{C} \in \boldsymbol{R}^{1\times 1\times 64}$
(4) The channel-wise attention descriptor,$\boldsymbol{C}$ which acts as a weight function to enhance the contributions of the significant and salient channels, is multiplied to the road marking feature map input, $\boldsymbol{I}_{\mathrm{FCA}}$ in a channel-wise manner to finally output the road marking feature map, denoted as $\boldsymbol{O}_{\mathrm{FCA}} \in \boldsymbol{R}^{H\times W\times 64\times 16}$ by

$$\overline{\boldsymbol{U}}_j^i = c_i \cdot \boldsymbol{U}_j^i \tag{5}$$

where c_i denotes the i-th element of $\boldsymbol{C}$; $\overline{\boldsymbol{U}}_j^i$ is the recalibrated j-th capsule in the i-th channel of $\boldsymbol{O}_{\mathrm{FCA}}$.

(2) **Class region attention (CRA) module**

The CRA module aims to highlight the spatial features by considering the impacts from all the other positions on road markings. Fig. 5 shows the architecture of the CRA module. As shown in Fig. 5, $I_{CRA} \in R^{H\times W\times 64\times 16}$ denotes a road marking feature input which is also composed of 64 16-dimensional capsule features with the size of $H \times W$ pixels. Concretely,

(1) Two 1 × 1 capsule convolution operations are applied to I_{CRA} to generate two one-dimensional capsule feature maps, $B \in R^{H\times W\times 64}$ and $\boldsymbol{D} \in \boldsymbol{R}^{H\times W\times 64}$, respectively, containing 64 channels with the identical size of the input road marking feature map.
(2) Feature maps $\boldsymbol{B}$ and $\boldsymbol{D}$ are reshaped to obtain feature matrices, $B_1 \in R^{N\times 64}$ and $\boldsymbol{D}_1 \in \boldsymbol{R}^{N\times 64}$, respectively, where $N = H \times W$. Then, the two resultant feature maps, $\boldsymbol{B}_1$ and $\boldsymbol{D}_1$, are computed by a matrix multiplication operation to generate a class region attention matrix, $\boldsymbol{S} \in \boldsymbol{R}^{N\times N}$. This can be obtained by

$$S(k,l) = \frac{exp\left(\sum_{m=1}^{64} B_1(k,m)D_1(m,l)\right)}{\sum_{n=1}^{N} exp\left(\sum_{m=1}^{64} B_1(n,m)D_1(m,l)\right)} \tag{6}$$

where $S(k,l)$ is the element at the k-th row and l-th column of $\boldsymbol{S}$. $B_1(k,m)$ and $D_1(m,l)$ are the element at the k-th row and m-th column of $\boldsymbol{B}_1$, and the element at the m-th row and l-th column of $\boldsymbol{D}_1$, respectively.

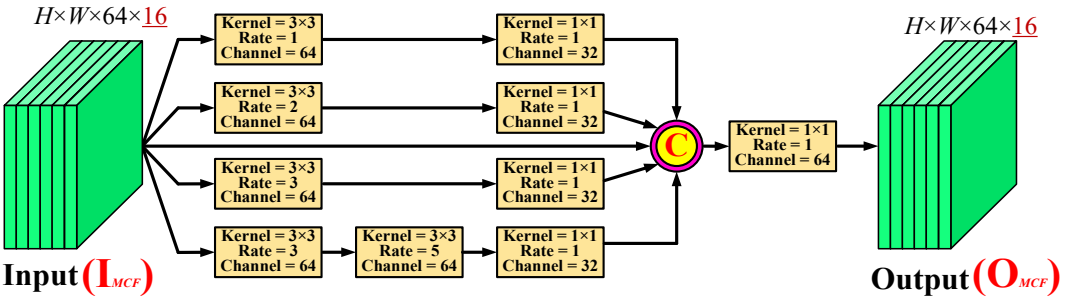


Fig. 3. Architecture of the MCF descriptor.

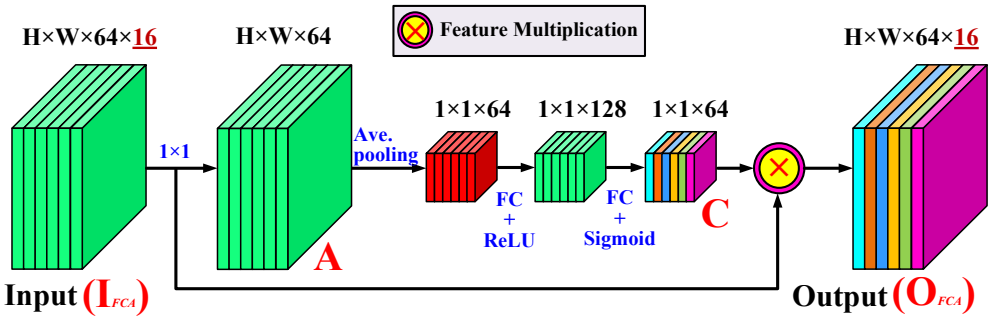


Fig. 4. Architecture of the FCA module.

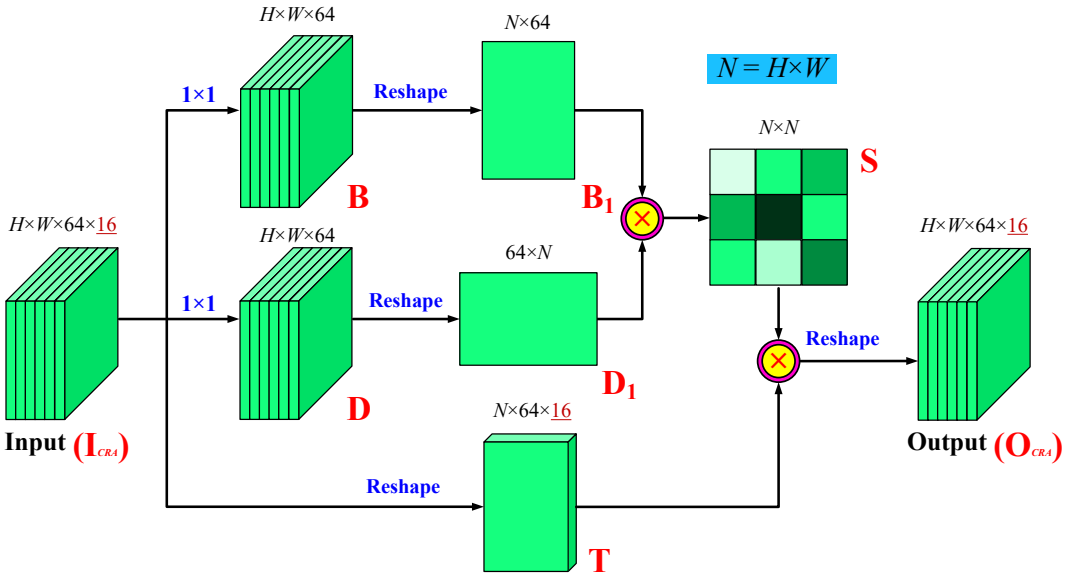


Fig. 5. Architecture of the CRA module.

(3) Afterwards, the road marking feature input, $\boldsymbol{I}_{\mathrm{CRA}}$, is reshaped to obtain a capsule feature matrix, $T \in R^{N\times 64\times 16}$, and a matrix multiplication operation is performed by multiplying it with the class region attention matrix, $\boldsymbol{S}$, to produce a recalibrated capsule feature matrix, which is finally reshaped to output a class region highlighted feature map, denoted as $\boldsymbol{O}_{\mathrm{CRA}} \in \boldsymbol{R}^{H\times W\times 64\times 16}$.

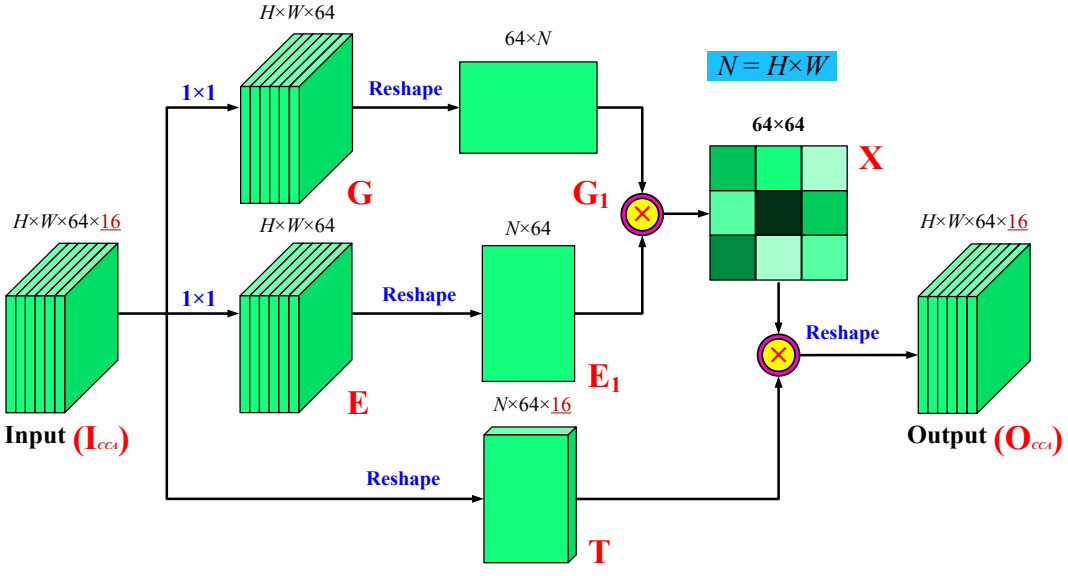


Fig. 6. Architecture of the CCA module.

(3) **Class channel attention (CCA) module**

The CCA module is designed to effectively emphasize the channel-wise features tightly related to the specific class features by taking into consideration the impacts from all the other channels on road markings. Fig. 6 shows the architecture of the CCA module. Similarly, a road marking feature input, $\boldsymbol{I}_{\mathrm{CCA}} \in \boldsymbol{R}^{H\times W\times 64\times 16}$, consists of 64 16-dimensional capsule features with the size of $H \times W$ pixels.

(1) Two 1 × 1capsule convolution operations are performed on $\boldsymbol{I}_{\mathrm{CCA}}$ to obtain two one-dimensional capsule feature maps, $\boldsymbol{G} \in \boldsymbol{R}^{H\times W\times 64}$ and $\boldsymbol{E} \in \boldsymbol{R}^{H\times W\times 64}$, respectively, containing 64 channels with the identical size of $\boldsymbol{I}_{\mathrm{CCA}}$.
(2) Feature maps **G** and **E** are reshaped to feature matrices, $\boldsymbol{G}_1 \in \boldsymbol{R}^{N\times 64}$ and $\boldsymbol{E}_1 \in \boldsymbol{R}^{N\times 64}$, respectively. A matrix multiplication operation is then performed between feature matrices $\boldsymbol{G}_1$ and $\boldsymbol{E}_1$, followed by a softmax function to construct a class channel attention matrix,$\boldsymbol{X} \in \boldsymbol{R}^{64\times 64}$. This is achieved by the following equation:

$$X(k,l) = \frac{exp\left(\sum_{m=1}^{64} G_1(k,m)E_1(m,l)\right)}{\sum_{n=1}^{64} exp\left(\sum_{m=1}^{N} G_1(n,m)E_1(m,l)\right)} \tag{7}$$

where $X(k,l)$ is the element at the k-th row and l-th column of the X, which measures the impact of channel k on the channel l in the input feature map. $G_1(k,m)$ and $E_1(m,l)$ are the element at the k-th row and m-th column of $\boldsymbol{G}_1$, and the element at the m-th row and l-th column of $\boldsymbol{E}_1$, respectively.

(3) Afterwards, the road marking feature input, $\boldsymbol{I}_{\mathrm{CCA}}$, is reshaped to produce a capsule feature matrix, $\boldsymbol{T} \in \boldsymbol{R}^{N\times 64\times 16}$, and then a matrix multiplication operation is performed by multiplying $\boldsymbol{T}$ with $\boldsymbol{X}$ to recalibrate the capsule feature matrix, which is finally reshaped to output a class channel emphasized feature map, denoted as $O_{CCA} \in R^{H\times W\times 64\times 16}$.

4. Results and discussion

4.1. Dataset and experimental setup

4.1.1. Dataset

In year 2020, we constructed a large UAV image dataset for road marking extraction tasks. We named this dataset as the RMS2020. This original UAV images were captured by a DJI Phantom 4 Pro system, and covered five different zones, a total of 10 square kilometers, in urban, suburban, and rural areas, Nanjing, Jiangsu province, China. The collected UAV images had a GSD of about 0.2 m, and then were processed to generate the RMS2020 dataset, containing about 20,000 images with the image size of 800 × 800 pixels. The RMS2020 dataset contains remarkably challenging road marking images. The images in the RMS2020 dataset are characterized by different spatial distributions and geometric topologies, varying intensities and sizes, diverse shapes and environmental conditions, and even different image qualities. In the RMS2020 dataset, 12,000 images were used as the training subset (60%), 1000 images as the validation subset (5%), and 7000 images as the test subset (35%), respectively, for our road marking extraction tests.

4.1.2. Model training & testing

The ACapsFPN was trained and performed on a cloud computing platform equipped with ten 16-GB GPUs, a 16-core CPU, and a 128-GB memory. Before training, a normal Gaussian distribution with the 0.01 standard deviation was used for drawing parameters to randomly initialize all layers of the ACapsFPN. The training images were organized into batches and fed into the ten GPUs to construct the ACapsFPN. The ACapsFPN was trained for 1500 epochs, each of which contained two images per GPU. We trained the model with 0.001 learning rate for the first 1200 epochs and 0.0001 learning rate for the rest 300 epochs.

Considering the varied orientation characteristics of road markings in the elevated-view UAV images, the training set was not directly applied to train the proposed ACapsFPN. At the training stage, to effectively construct a high-performance road marking extraction model, data augmentation was carried out to enlarge the training set. Specifically, we generated a horizontal mirror image for each training image in the horizontal direction. Both the horizontal mirror image and its corresponding original image were then clockwise rotated in four directions at a 90 degrees' angle interval. In such a way, each training image in the training set was converted into eight images. Correspondingly, the road marking label map was also transformed to generate the ground truths. Therefore, after data augmentation, the final training set contained 96,000 images, which was eight times in size of the initial training set, to train the ACapsFPN. To quantitatively evaluate the robustness and superiority of the proposed ACapsFPN, four metrics, i.e., *precision*, *recall*, *intersection-over-union* (*IoU*), and F_1-*score*, were used by comparing the extracted road marking results with the labelled maps.

4.2. Road-marking extraction

To evaluate the road marking extraction performance of our ACapsFPN, we applied it to the RMS2020 dataset. As shown in Table 1, the proposed ACapsFPN obtained a *precision* of 0.7366, a *recall* of 0.7513, an *IoU* of 0.5922, and an F_1-*score* of 0.7439, respectively, for road markings. The extraction results were quite promising when handling the complicated and challenging RMS2020 dataset. Fig. 7 illustrates a small group of representative road marking extraction results generated by the ACapsFPN. The extracted road markings were colored in red. Although the great variations of the road markings in spatial sizes, intensity appearances, geometric topologies, and complicated scenarios (see Fig. 7), the ACapsFPN differentiated well the road markings from the surrounding environments with a fairly small proportion of false alarms and missing detections. Concretely, due to occlusions caused by vehicles, pedestrians, and other objects, as well as varying illumination conditions, the road markings exhibited with different patterns and intensity incompleteness in the UAV images. In different areas, the geometric topologies of the road markings varied greatly in shapes and sizes, and the road marking distribution patterns and types were quite different from one area to another area. For instance, the road markings on the main road areas were usually simply distributed with generally lane markings (see Fig. 7 (a), (d), and (k)). In contrast, the road markings at the crosswalk areas were densely distributed with different types of road markings (see Fig. 7 (c), (f), (g), and (j)). In addition, some road markings were partially occluded by the on-road or overhead objects, which changed the completeness and the geometric topologies of the road markings (see Fig. 7 (g) and (l)). The

Table 1
Extracted Road Marking Results Obtained by different Methods.

Networks	Quantitative Evaluation			
	precision	*recall*	*IoU*	F_1-*score*
CapsFPN-1	0.7165	0.7346	0.5692	0.7254
CapsFPN-2	0.7138	0.7315	0.5656	0.7225
CapsFPN-3	0.6874	0.7011	0.5316	0.6942
Aerial LaneNet (Azimi et al., 2019)	0.6621	0.6795	0.5045	0.6707
Modified U-Net (Wen et al., 2019)	0.6388	0.6597	0.4805	0.6491
Deeplabv3 (Chen et al., 2017)	0.7032	0.7215	0.5531	0.7122
PSPNet (Zhao et al., 2017)	0.7026	0.7207	0.5522	0.7115
U-Shaped Capsule Network (Ma et al., 2021a,b)	0.6681	0.6837	0.5104	0.6758
SA-CapsFPN (Yu et al., 2021)	0.7143	0.7322	0.5663	0.7231
HRNet (Wang et al., 2021)	0.7106	0.7287	0.5619	0.7195
DFPN (Chen et al., 2021)	0.7053	0.7231	0.5553	0.7141
ACapsFPN	0.7366	0.7513	0.5922	0.7439

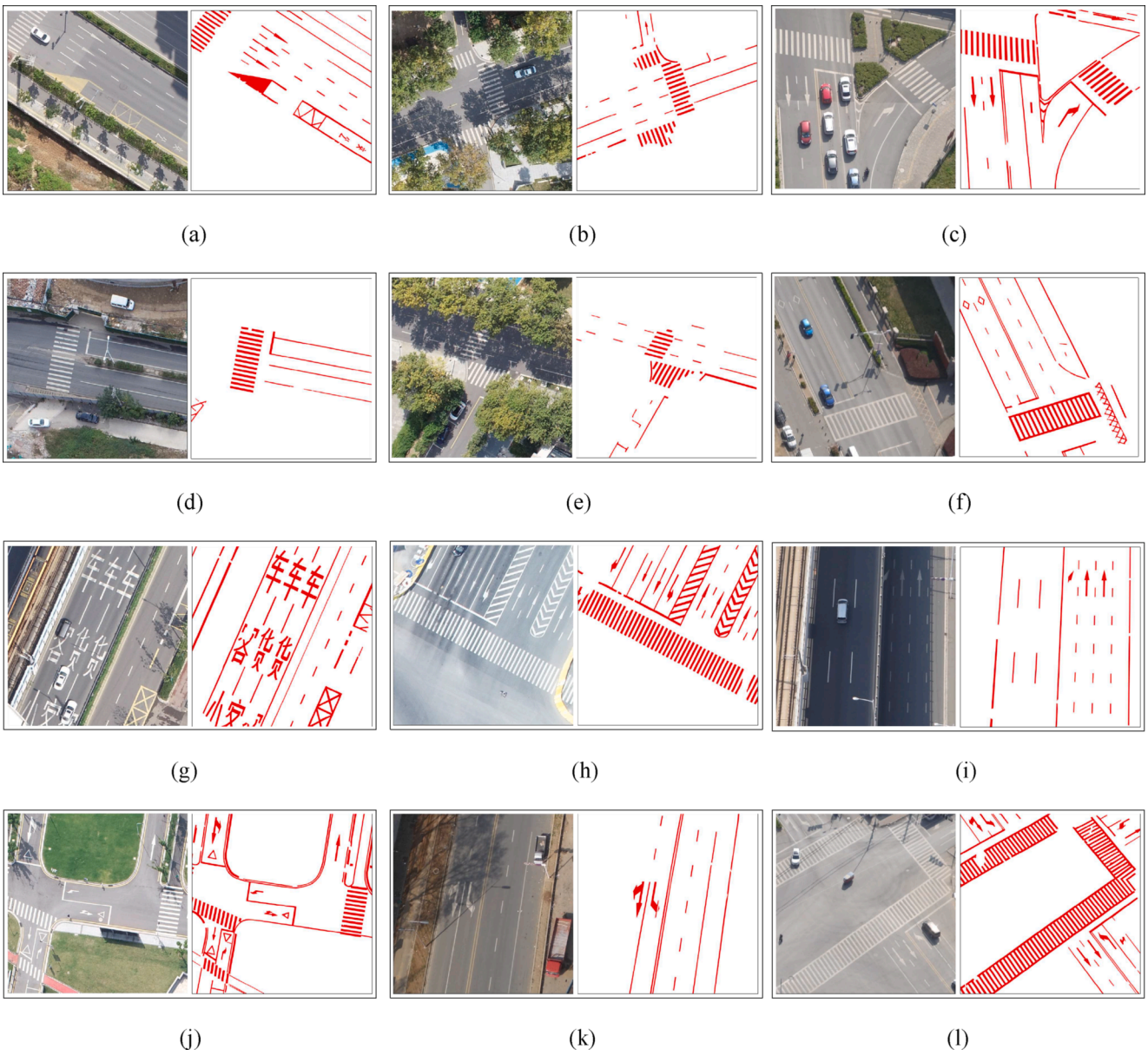


Fig. 7. A close view of the road marking extraction results obtained by the ACapsFPN.

shadows cast on the road markings from the nearby objects also influenced the correct identification of the complete road marking regions (see Fig. 7 (b), (e), and (k)). Moreover, in some images, the contrasts between the road markings and their road environments were extremely low or quite fuzzy (see Fig. 7 (i) and (l)), which brought great challenges to accurately segment these road markings from the UAV images. Fortunately, for the challenging RMS2020 dataset, by adopting a feature pyramid architecture by fusing the capsule features at different levels and different scales, the ACapsFPN enhances the feature representation capability and the localization accuracy. Next, embedded with the MCF descriptor and the FCA module, the ACapsFPN is capable of using different scales of contextual properties and highlighting the salient and informative features, thereby further improving the capability of road marking feature representation. Additionally, designed with the CRA and CCA modules, the ACapsFPN concentrates on the spatial features related to the road markings and the feature channels tightly linking with the road markings. Therefore, the proposed ACapsFPN performed promisingly in processing the UAV images containing road markings of different challenging scenarios and provided an acceptable road marking extraction result. However, as shown in Fig. 7, some objects (e.g., traffic poles or some doodling lines) were falsely detected as road markings due to their similar textural and geometric properties to the road markings. Moreover, caused by heavy shadows, severe occlusions, and abrasions, some road markings were partially hidden into the background or disjointed into several parts. Thus, the completeness of these road markings were failed to be maintained.

Aiming at further assessing the applicability and transferability of the ACapsPFN, we also collected a new set of UAV images. To facilitate testing, all the images were cropped into patches of 800 × 800 pixels with an overlap size of 200 pixels. The road marking extraction results of each patch were fused to generate the final extraction results of the corresponding UAV image. As shown in Fig. 8, we overlaid the extracted road markings onto the raw UAV images. We found that the proposed ACapsPFN was capable of accurately extracting most road markings shadowed by trees and buildings, or varying illumination conditions. This is because multi-scale feature abstraction and fusion contributes to the inference of road markings. Of course, our ACapsPFN still faced the challenges of the severely-eroded and largely-occluded road markings.

Fig. 8. Road marking extraction results obtained by the ACapsFPN.

The road markings occluded by objects, such as vehicles and trees, were not continuously delineated due to no road making data appeared on the images. There were some misclassified road markings because of the high spectral similarity of road markings and linear objects, such as the arms of light poles. Note that lane markings were almost completely extracted (see Fig. 8). Precisely and accurately identifying and localizing lane road markings is crucial for road asset management and budget allocation for road maintenance. Overall, the road marking extraction performance showed that the ACapsFPN had a promising generalization capability and behaved accurately and competitively in handling varied-shape road markings of different self-conditions in varying scenarios. Moreover, to evaluate the computational performance of the proposed ACapsPFN, the processing time was also recorded on the test datasets at the road marking extraction stage. On average, the ACapsPFN achieved a processing speed of about 16 image patches per second on a GPU.

4.3. Ablation analysis

As ablation experiments, we further demonstrated the competitive performance achieved by the MCF descriptor and the ternary feature attention modules. Specifically, the MCF descriptor functioned to collect and aggregate multi-scale contextual road marking information with different-size receptive fields. The FCA module aimed to emphasize the salient and information-rich features and suppress the useless ones to strengthen the capability of its road marking feature representation. The CRA module focused on the spatial features related to road markings, and the CCA module highlighted the feature channels tightly associated with road markings, which provided a high-quality road marking feature representation. All of these modules improved the accuracy of pixel-wise road marking extraction. To conduct the ablation experiments, we constructed three networks on the basis of the proposed ACapsFPN. Concretely, first, we removed the CRA and the CCA modules (integrating only the MCA and the FCA modules) from the ACapsFPN. We named the resultant network as the CapsFPN-1. Then, we removed all the MCA and the FCA modules (integrating only the CRA and the CCA modules) at each stage of the ACapsFPN. We named the resultant network as the CapsFPN-2. Finally, we removed the MCF descriptor and the ternary feature attention modules from the ACapsFPN, resulting in a pure capsule feature pyramid network without any feature augmentation and feature attention mechanisms. We named the resultant network as the CapsFPN-3. We trained the three networks with the same training and validation sets, as well as the same data augmentation strategy.

Afterwards, these three constructed networks were applied to the test set to evaluate their performances on road marking extraction. Table 1 lists the quantitative results obtained by the three networks. Obviously, without the feature augmentation and the ternary feature attention modules, the road marking extraction accuracy of the CapsFPN-3 was significantly degraded. The accuracy degradation was mainly due to the following factors: the very small-size road markings, the road markings occluded severely by the nearby objects, the worn-out road markings, the road markings covered with heavy shadows, or the road markings showing quite low contrasts with their road surroundings. Thus, the CapsFPN-3 behaved less effectively in processing such challenging road marking scenarios. However, the overall performance was still acceptable. This results from the use of the capsule neurons to characterize high-order entity features and the design of the feature pyramid network architecture to fuse multiscale features. In contrast, with the embedding of the feature augmentation modules and the ternary feature attention modules for, respectively, aggregating multi-scale contextual properties without losing feature resolutions and details and highlighting the contributions of the informative and class-specific features, the road marking extraction performances were dramatically improved by the CapsFPN-1 and CapsFPN-2. Fig. 9 shows the comparative results. Fig. 9 (a) and (b) shows an original image and its corresponding ground truth. Fig. 9 (c) – (f) demonstrates the results obtained by ACapsFPN, CapsFPN-1, CapsFPN-2, and CapsFPN-3, respectively. We found that our ACapsFPN and CapsFPN-1 partially extracted some road markings occluded by trees, whereas CapsFPN-2 and CapsFPN-3 failed to extract them, as shown in the green boxes. This indicated that the MCF descriptor contributes to include more contextual information by using different atrous rates. Especially, CapsFPN-3 failed to extract the road markings of a bus stop region largely occluded by tree shadows. Through ablation experiments, we concluded that the MCF descriptor and the ternary feature attention modules (i.e., the FCA, CRA, and CCA modules) contributed positively and powerfully to the upgradation of the road marking extraction accuracy. Therefore, due to the cooperation of these modules, the proposed ACapsFPN showed advantageous performance in handling the UAV images containing road marking instances of different geometric topologies and distributions, varying intensities and sizes, and diverse shapes and environmental conditions.

4.4. Comparative tests

To further evaluate the robustness of our ACapsFPN in road marking extraction tasks, we compared it with recently-presented road marking extraction methods and semantic segmentation methods, i.e., Aerial LaneNet (Azimi et al., 2019), Modified U-Net (Wen et al., 2019), Deeplabv3 (Chen et al., 2017), PSPNet (Zhao et al., 2017), U-Shaped Capsule Network (Ma et al., 2021a,b), SA-CapsFPN (Yu et al., 2021), HRNet (Wang et al., 2021), and DFPN (Chen et al., 2021). Specifically, Aerial LaneNet used an FCN architecture. PSPNet is a multi-scale scene parsing network, which uses pyramid pooling (SPP) module for image segmentation. DeepLabv3 employs an atrous spatial pyramid pooling (ASPP) module to extract the contextual features at different scales. U-Shaped Capsule Network and SA-CapsFPN were based on capsule networks by adopting a U-Net architecture and an FPN architecture,

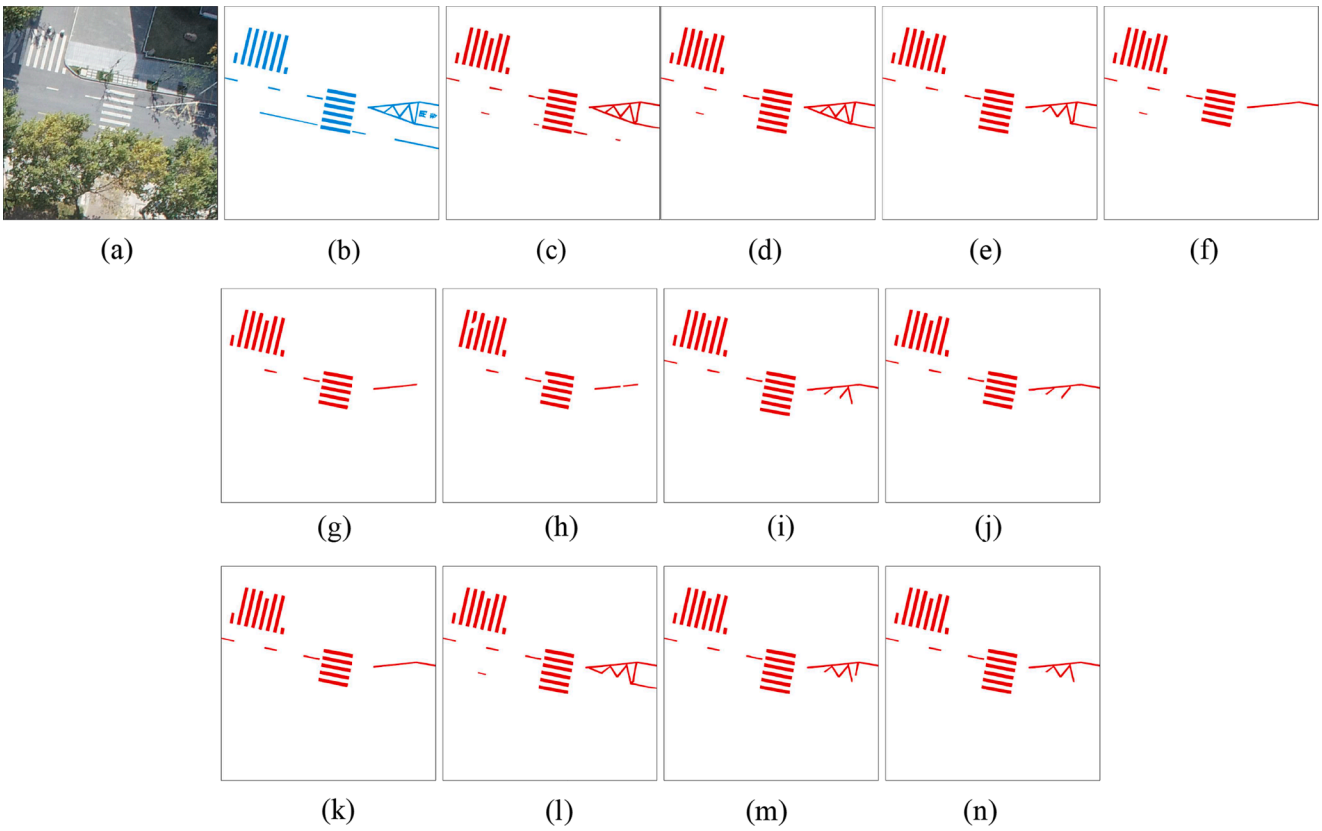


Fig. 9. Road marking extraction results obtained by the comparative methods. (a) a raw image, (b) ground truth, (c) ACapsFPN, (d) CapsFPN-1, (e) CapsFPN-2, (f) CapsFPN-3, (g) Aerial LaneNet, (h) Modified U-Net, (i) Deeplabv3, (j) PSPNet, (k) U-Shaped Capsule, (l) SA-CapsFPN, (m) HRNet, and (n) DFPN.

respectively. Moreover, besides SA-CapsFPN and our method, DFPN also used an FPN architecture. Specifically, multilevel features were considered and properly fused in some methods for improving the pixel-wise road marking extraction accuracy, and attention mechanisms were also utilized in some methods to further enhance the feature representation capability. For fair comparisons, the same training, validation, and test sets, as well as the same data augmentation strategy were used to train and evaluate the comparative models. Table 1 shows the quantitative road marking results obtained by these models.

As shown in Table 1, SA-CapsFPN achieved the better overall road marking extraction accuracies. In contrast, Modified U-Net, Aerial LaneNet, and U-Shaped Capsule Network methods achieved less effectively than the other methods. The rest methods showed similar road marking extraction performances. For SA-CapsFPN, DFPN, and ACapsFPN, an FPN architecture was used for extracting and fusing multi-level and multi-scale features. For Modified U-Net, as a special encoder-decoder network, road marking features were extracted by a contraction path, the details were restored via a corresponding expansion path, and then the two paths were connected by skip connections to enhance feature information. However, for the road markings in the RMS2020 dataset, some road markings had extremely challenging self-conditions and complicated surrounding environments. The simple architectures, such as U-Net and FCN, performed less promisingly in correctly segmenting road markings. For the U-Shaped Capsule Network, capsule networks contributed to the improvement of road marking extraction due to its capsule based feature representations. For SA-CapsFPN, the MCF descriptor and the FCA module were integrated into the FPN architecture to exploit multi-scale contextual information and emphasize useful channel features, improving the capability of feature representation.

Figure 9 shows the road marking results obtained by different methods. Visual inspection also demonstrated that ACapsFPN outperformed other comparative methods. For example, as shown in the black boxes, ACapsFPN and SA-CapsFPN were capable of extracting the most road markings of the bus stop region, while Aerial LaneNet, Modified U-Net, and U-Shaped Capsule failed to deal with these road markings occluded by tree shadows. For the lane markings in the green boxes, all the methods faced this challenge that the road markings were completely occluded by the trees. Because of this kind of occlusion caused by trees, buildings, or image perspectives, no lane marking data can be shown on the image, resulting in the failure of road marking extraction. Comparatively, our ACapsFPN was superior to the eight compared methods in road marking extraction because of the following reasons: (1) the capsule FPN architecture contributes to the extraction and fusion of multi-level high-order features, (2) the MCF descriptor helps effectively exploit multi-scale contextual information at a high-resolution perspective, and the FCA module emphasizes the important and salient channel features, (3) the CRA module and the CCA module highlight the spatial features connected with the road markings and the feature channels tightly related to the road markings. To sum up, comparative analysis demonstrated that the ACapsFPN provided an effective and promising road marking extraction method by using high-resolution UAV images.

5. Conclusions

This paper presented a novel attentive capsule feature pyramid network, named ACapsFPN, to accurately and precisely segment road markings. In the ACapsFPN, a deep capsule FPN architecture, which was capable of extracting and fusing multi-level capsule features at different scales, was employed to output high-quality and task-aware feature semantics for generating a highly accurate road marking map. By integrating the MCF descriptor and the FCA module into each stage of the hierarchical encoder, the ACapsFPN can rapidly exploit contextual

properties at multiple scales with a high-resolution view, and enhance the channel-wise significant and salient features, which functioned positively to further improve the capability of feature abstraction. In addition, by designing the CRA module and the CCA module over the fused features, the ACapsFPN concentrated on the spatial features of road markings and highlighted the feature channels tightly related to the road markings, which effectively suppressed the influence of the background and provided a high-quality road marking feature encoding. We evaluated the ACapsFPN on the RMS2020, a large-volume high-resolution UAV image dataset, and gained an excellent performance in extracting road markings with different spatial distributions and geometric topologies, varying intensities and sizes, diverse shapes and environmental conditions, and even different image qualities. Quantitative assessments demonstrated that a *precision* of 0.7366, a *recall* of 0.7513, an *IoU* of 0.5922, and an F_1*-score* of 0.7439, respectively, were achieved in segmenting road markings from the high-resolution UAV images. Comparative experiments with eight state-of-the-art methods also convinced the competitive and advantageous performance of the ACapsFPN in road marking extraction tasks.

CRediT authorship contribution statement

Haiyan Guan: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Project administration, Funding acquisition. **Xiangda Lei:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Yongtao Yu:** Validation, Investigation, Resources, Visualization, Supervision. **Haohao Zhao:** Data curation, Writing – review & editing, Funding acquisition. **Daifeng Peng:** Data curation, Writing – review & editing, Funding acquisition. **José Marcato Junior:** Writing – review & editing. **Jonathan Li:** Resources, Supervision, Writing – review & editing.

Funding

This research was supported by the National Natural Science Foundation of China [grant numbers 41971414, 41801239, 62001175], Natural Science Foundation of Jiangsu Province [No. BK20211365], and Natural Science Foundation of Fujian Province [No.2021J01081].

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.

References

Azimi, S.M., Fischer, P., Korner, M., Reinartz, P., 2019. Aerial LaneNet: lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 57 (5), 2920–2938.

de Paula, M.B., Jung, C.R., 2015. Automatic detection and classification of road lane markings using onboard vehicular cameras. IEEE Trans. Intell. Transport. Syst. 16 (6), 3160–3169.

Chen, L., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. CoRR, vol. abs/1706.05587, 2017. [Online]. Available: http://arxiv.org/abs/1706.05587.

Chen, S., Zhang, Z., Zhong, R., Zhang, L., Ma, H., Liu, L., 2021. A dense feature pyramid network-based deep learning model for road marking instance segmentation using MLS point clouds. IEEE Trans. Geosci. Remote Sens. 59 (1), 784–800.

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J., 2018. Cascaded pyramid network for multi-person pose estimation. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2018, 7103–7112.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2014, 580–587.

Girshick, R., 2015. R-CNN Fast. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2015, 1440–1448.

Gopalan, R., Hong, T., Shneier, M., Chellappa, R., 2012. A learning approach towards detection and tracking of lane markings. IEEE Trans. Intell. Transp. Syst. 13 (3), 1088–1098.

Grabner, H., Nguyen, T.T., Gruber, B., Bischof, H., 2008. On-line boosting based car detection from aerial images. ISPRS J. Photogramm. Remote Sens. 63 (3), 382–396.

Greenhalgh, J., Mirmehdi, M., 2015. Detection and recognition of painted road surface markings. Proc. Int. Conf. Pattern Recognit. Appl. Methods 2015, 130–138.

Gupta, A., Choudhary, A., 2018. A framework for camera-based real-time lane and road surface marking detection and recognition. IEEE Trans. Intell. Vehicles 3 (4), 476–485.

Han, S., Han, Y., Hahn, H., 2009. Vehicle detection method using Haar-like feature on real time system. World Acad. Sci. Eng. Technol. 59, 455–459.

He, B., Ai, R., Yan, Y., Lang, X., 2016. Accurate and robust lane detection based on dual-view convolutional neutral network. In: Proceedings of the IEEE Intell. Vehicles Symp., Gothenburg, Sweden, Jun. 2016, pp. 1041–1046.

Hoang, T.M., Nam, S.H., Park, K.R., 2019. Enhanced detection and recognition of road markings based on adaptive region of interest and deep learning. IEEE Access 7, 109817–109832.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2018, 7132–7141.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2017, 4700–4708.

Jung, S., Youn, J., Sull, S., 2016. Efficient lane detection based on spatiotemporal images. IEEE Trans. Intell. Transp. Syst. 17 (1), 289–295.

Kim, Z.W., 2008. Robust lane detection and tracking in challenging scenarios. IEEE Trans. Intell. Transport. Syst. 9 (1), 16–26.

Lee, C., Moon, J.-H., 2018. Robust lane detection and tracking for real-time applications. IEEE Trans. Intell. Transp. Syst. 19 (12), 4043–4048.

Lee, S., Kim, J., Yoon, J. S., Shin, S., Bailo, O., Kim, N., Lee, T.-H., Hong, H. S., Han, S.-H., Kweon, I. S., 2017. VPGNet: Vanishing point guided network for lane and road marking detection and recognition. In: Proc. IEEE Int. Conf. Comput. Vis., Venice, Italy, Oct. 2017, pp. 1965–1973.

Li, J., Mei, X., Prokhorov, D., Tao, D., 2017. Deep neural network for structural prediction and lane detection in traffic scene. IEEE Trans. Neural Netw. Learn. Syst. 28 (3), 690–703.

Li, M., Li, Y., Jiang, M., 2018. Lane detection based on connection of various feature extraction methods. Adv. Multimedia 2018, 1–13.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2117–2125.

Lyu, Y.e., Vosselman, G., Xia, G.-S., Yilmaz, A., Yang, M.Y., 2020. UAVid: A semantic segmentation dataset for UAV imagery. ISPRS J. Photogramm. Remote Sens. 165, 108–119.

Ma, L., Li, Y., Li, J., Yu, Y., Junior, J.M., Goncalves, W.N., Chapman, M.A., 2021a. Capsule-based networks for road marking extraction and classification from mobile LiDAR point clouds. IEEE Trans. Intell. Transp. Syst. 22 (4), 1981–1995.

Ma, X., Zhong, H., Li, Y., Ma, J., Cui, Z., Wang, Y., 2021b. Forecasting transportation network speed using deep capsule networks with nested LSTM models. IEEE Trans. Intell. Transp. Syst. 22 (8), 4813–4824. https://doi.org/10.1109/TITS.2020.2984813.

Mathibela, B., Newman, P., Posner, I., 2015. Reading the road: road marking classification and interpretation. IEEE Trans. Intell. Transport. Syst. 16 (4), 2072–2081.

McCall, J.C., Trivedi, M.M., 2006. Video-based lane estimation and tracking for driver assistance: Survey, system, and evaluation. IEEE Trans. Intell. Transp. Syst. 7 (1), 20–37.

Niu, J., Lu, J., Xu, M., Lv, P., Zhao, X., 2016. Robust lane detection using two-stage feature extraction with curve fitting. Pattern Recognit. 59, 225–233.

Ozgunalp, U., Fan, R., Ai, X., Dahnoun, N., 2017. Multiple lane detection algorithm based on novel dense vanishing point estimation. IEEE Trans. Intell. Transport. Syst. 18 (3), 621–632.

Paoletti, M.E., Haut, J.M., Fernandez-Beltran, R., Plaza, J., Plaza, A., Li, J., Pla, F., 2019. Capsule networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 57 (4), 2145–2160.

Prakash, T., Comandur, B., Chang, T., Elfiky, N., Kak, A., 2015. A generic road-following framework for detecting markings and objects in satellite imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8 (10), 4729–4741.

Sabour, S., Frosst, N., Hinton, G.E., 2017. Dynamic routing between capsules. In: Proc. 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–10 Dec. 2017, pp. 1–11.

Shamsolmoali, P., Chanussot, J., Zareapoor, M., Zhou, H., Yang, J., 2021a. Multipatch feature pyramid network for weakly supervised object detection in optical remote sensing images. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2021.3106442.

Shamsolmoali, P., Zareapoor, M., Chanussot, J., Zhou, H., Yang, J., 2021b. Rotation equivariant feature image pyramid network for object detection in optical remote sensing imagery. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2021.3112481.

Shamsolmoali, P., Zareapoor, M., Zhou, H., Wang, R., Yang, J., 2021c. Road segmentation for remote sensing images using adversarial spatial pyramid networks. IEEE Trans. Geosci. Remote Sens. 59 (6), 4673–4688.

Son, J., Yoo, H., Kim, S., Sohn, K., 2015. Real-time illumination invariant lane detection for lane departure warning system. Expert Syst. Appl. 42 (4), 1816–1824.

Tian, Y., Gelernter, J., Wang, X., Chen, W., Gao, J., Zhang, Y., Li, X., 2018. Lane marking detection via deep convolutional neural network. Neurocomput. 280, 46–55.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B., 2021. Deep high-resolution representation learning for

visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43 (10), 3349–3364. https://doi.org/10.1109/TPAMI.2020.2983686.
Wen, C., Sun, X., Li, J., Wang, C., Guo, Y., Habib, A., 2019. A deep learning framework for road marking extraction, classification and completion from mobile laser scanning point clouds. ISPRS J. Photogramm. Remote Sens. 147, 178–192.
Xiao, D., Yang, X., Li, J., Islam, M., 2020. Attention deep neural network for lane marking detection. Knowl-Based Syst. 194 (105584), 1–10.
Xu, X., Yu, T., Hu, X., Ng, W.W.Y., Heng, P.-A., 2021. SALMNet: a structure-aware lane marking detection network. IEEE Trans. Intell. Transport. Syst. 22 (8), 4986–4997.
Ye, X., Hong, D., Chen, H., Hsiao, P., Fu, L., 2020. A two-stage real-time YOLOv2-based road marking detector with lightweight spatial transformation-invariant classification. Image Vis. Comput. 102 (103978), 1–11.
Yu, Y., Yao, Y., Guan, H., Li, D., Liu, Z., Wang, L., Yu, C., Xiao, S., Wang, W., Chang, L.v., 2021. A self-attention capsule feature pyramid network for water body extraction from remote sensing imagery. Int. J. Remote Sens. 42 (5), 1801–1822.
Zhang, A., Wang, K.C.P., Yang, E., Qiang Li, J., Chen, C., Qiu, Y., 2018. Pavement lane marking detection using matched filter. Measurement 130, 105–117.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2017, 2881–2890.
Zhu, Q., Huang, S., Hu, H., Li, H., Chen, M., Zhong, R., 2021. Depth-enhanced feature pyramid network for occlusion-aware verification of buildings from oblique images. ISPRS J. Photogramm. Remote Sens. 174, 105–116.