RGB-T Nested Semantic Segmentation Network Fusing Deep Differential Features: Postprint

Authors: Yuan Haobin, Zhao Tao, Zhong Yuzhong

Date: 2022-05-10T11:22:58+00:00

Abstract

To address the suboptimal segmentation performance of existing visible-infrared (RGB-T) image semantic segmentation models, this paper proposes a nested segmentation network based on deep difference feature complementary fusion. Specifically, the encoder and decoder components are interconnected via multi-level dense intermediate paths to form a nested architecture, wherein the decoder leverages multi-level pathways to densely reuse shallow and deep features from the encoder for multi-scale feature exploitation. Additionally, multi-modal deep features are enhanced in their semantic representation capability through a feature difference fusion strategy. Experimental results demonstrate that the proposed network achieves 65.8% mean accuracy and 54.7% mean Intersection over Union (mIoU) on the MFNet dataset, exhibiting superior segmentation performance compared to other state-of-the-art RGB-T segmentation models.

Full Text

Nested Semantic Segmentation Network Fusing Deep Difference Features

Yuan Haobin, Zhao Tao, Zhong Yuzhong
(College of Electrical Engineering, Sichuan University, Chengdu 610065, China)

Abstract: Existing visible-infrared (RGB-T) image semantic segmentation models suffer from limited segmentation performance. To address this issue, we propose a nested segmentation network based on complementary fusion of deep difference features. Specifically, the encoding and decoding components of the network are connected through multi-level dense intermediate pathways to form a nested architecture, enabling dense multi-scale feature reuse for the decoder through hierarchical pathways from the encoder. Additionally, multi-modal deep features are enhanced in their semantic expressiveness through a feature differential fusion strategy. Experimental results demonstrate that the

proposed network achieves 65.8% mean accuracy and 54.7% mean Intersection over Union (mIoU) on the MFNet dataset, exhibiting superior segmentation capability compared to other state-of-the-art RGB-T segmentation models.

Keywords: RGB-T semantic segmentation; nested network; feature reuse; fusion strategy

0 Introduction

Semantic segmentation aims to assign pixel-level category labels to images, with broad applications in autonomous driving [1], medical analysis [2], and robot localization [3]. Due to limitations in visible light sensor imaging mechanisms [4], current mainstream RGB segmentation models inevitably experience performance degradation under conditions such as dense fog and low light [5]. Benefiting from infrared sensors' ability to capture thermal radiation information, infrared images can effectively compensate for missing information in RGB images under adverse conditions [6], making the fusion of these two modalities more robust for scene representation.

RGB-T semantic segmentation has attracted considerable research attention in recent years. MFNet [7] pioneered real-time RGB-T semantic segmentation for autonomous driving, drawing inspiration from the FuseNet architecture [8] with two symmetric low-parameter encoders and a single decoder. The final two encoder layers capture multi-scale features with larger receptive fields through micro-downsampling perception modules. RTFNet [9] utilizes ResNet [10] as the backbone for both encoders to integrate RGB and infrared information, with the decoder progressively restoring resolution and reconstructing features through two types of upsampling modules. Xu et al. [11] improved the encoder to a ResNet with dilated convolution operations to enhance small object detection, and designed a co-attention mechanism module to fuse extracted multi-modal features. Guo et al. [12] focused on multi-scale information utilization, proposing an auxiliary decoding module to receive features from all encoder levels, enabling more flexible context fusion through cross-scale feature transmission.

While these studies have contributed to RGB-T semantic segmentation at various levels, several challenges remain to be addressed. First, relying solely on deep features transmitted unidirectionally to sequentially connected decoder layers leads to loss of edge detail information due to encoding downsampling [9,11]. Although skip connections partially alleviate this by reusing same-scale encoder features at the decoder [7], the utilization of shallow and deep features remains insufficient. Moreover, encoders do not adequately account for feature modality differences between RGB and infrared images during fusion. For instance, in nighttime environments, infrared images contain information imperceptible to RGB images. Simple addition [9] or channel-wise concatenation [7] can, in certain cases, cause hedging effects on easily identifiable features, weakening the encoding response of dominant features—particularly impacting high-dimensional features more significantly. Meanwhile, co-attention fusion based on Softmax

operators [11] lacks learnable parameters for more comprehensive reuse of multi-level encoder features and mitigating modality differences in high-dimensional feature fusion.

To enable more comprehensive reuse of multi-level encoder features and reduce the impact of modality differences on high-dimensional feature fusion, this paper proposes an RGB-T nested semantic segmentation network that fuses deep difference features from RGB and infrared images. The contributions are:

a) **Dense reuse of encoder shallow and deep features.** The encoder and decoder are connected through multi-level intermediate pathways. Scale-diverse encoder features from different levels are integrated via stacking and fed to the decoder, enabling utilization of richer multi-scale feature information for semantic partitioning.

b) **Deep feature fusion strategy.** During deep feature fusion, considering the inherent differences between RGB and infrared images, we design a feature differential fusion strategy to extract complementary features from both modalities, achieving better information fusion and thereby enhancing the semantic representation capability of deep high-dimensional abstract features.

1 Nested Semantic Segmentation Network

The nested connection architecture was first proposed by Zhou et al. [13] for medical image segmentation tasks. Based on the observation that features at different levels exhibit varying sensitivity to objects of different sizes, they replaced the long skip connections in U-Net [14] with a nested combination of upsampling and short/long skip connections. Figure 1 illustrates the nested connection framework.

In nested structures, shallow and deep encoder features undergo dense concatenation and reuse in the channel dimension through upsampling and dense connections, enabling effective integration of features at various levels. Inspired by this, we introduce the nested structure into RGB-T semantic segmentation to construct a network that can fully integrate multi-scale feature information. As shown in Figure 2, the proposed segmentation model comprises two identically structured encoders and one decoder. The dual encoders on the left perform hierarchical downsampling to extract shallow and deep features, while the decoder on the right progressively reconstructs features. The encoding and decoding components are connected through densely connected multi-level intermediate transition layers, forming an overall nested architecture. Compared with existing RGB-T segmentation networks, the dense intermediate information flow channels effectively preserve semantic feature information at all levels.

1.2 Deep Difference Feature Complementary Fusion

The final dense feature unit transmits deep information through only a single channel, presenting a challenge during decoding reconstruction: deep networks capture limited gradient information for difficult targets such as small-scale objects. At this stage, RGB and infrared features exhibit higher-dimensional abstract semantics. Particularly under adverse lighting conditions, blind spot information carried by RGB images makes their deep features more difficult to learn. In such cases, incorporating infrared information should focus more on compensating for weak feature regions of both modalities. Given the imaging principle differences between RGB and infrared images, we propose a complementary fusion strategy based on feature differences constructed at the pixel level to enhance semantic expression of deep features.

As shown in Figure 3, the differential fusion module takes RGB and infrared feature maps as input. During RGB deep feature encoding, dual-modal features first undergo convolution operations to obtain compressed feature mapping matrices Q and K. These matrices are spatially unfolded and processed through the following operation to obtain a modality feature difference weight matrix:

$$W = 1 - \text{softmax}(Q^T K)$$

Feature maps manifest as numerical vector matrices at the pixel level. The multiplication of Q and K reflects the feature correlation between RGB and infrared features. The softmax normalization ensures the correlation matrix represents weight coefficients reflecting common features across global positions. Therefore, modality feature differences can be expressed through the complement to 1. Subsequently, the linear transformation matrix V of the RGB feature map is weighted with W to obtain complementary features of the RGB feature map:

$$\text{Feature}_{\text{com}}^{\text{RGB}} = V \cdot W$$

Similarly, complementary features of the infrared feature map $\text{Feature}_{\text{com}}^{\text{IR}}$ are obtained through the same process. Finally, the two complementary features are added to the input dual-modal features to achieve deep feature complementary fusion enhancement.

1.1 Multi-level Reuse of Shallow and Deep Features

Many RGB-T segmentation models adopt ResNet as the backbone. Considering that DenseNet [15] offers denser information propagation pathways with fewer parameters, this paper employs the DenseNet framework for the encoder backbone. To preserve more original spatial information and enhance structural uniformity within the encoder, the classification layer of DenseNet is removed, and an additional transition layer consistent with other transition layers is appended after the fourth dense block. Consequently, the encoder can be divided

into an initial convolution layer, a max pooling layer, and four dense feature units composed of dense blocks and transition layers. Dense blocks maintain feature map resolution, while the remaining components perform 2× downsampling. Considering that infrared images are single-channel grayscale, the input channel number of the initial convolution layer in the infrared encoder is modified to 1. For the first five downsampling processes, RGB and infrared information are fused through element-wise addition. For deep high-dimensional features extracted during the final downsampling stage, fusion is completed through the feature differential fusion strategy.

In the proposed model, fused features at each layer undergo multi-level backflow through upsampling and intermediate layers. The backflow features and output from previous fusion layers are densely stacked together and transmitted to the input of the corresponding-level reconstruction layer. Compared with using only long skip connections, the semantic gap between encoder and decoder layers can be alleviated through intermediate layers. As shown in Figure 1, the upsampling unit resembles a residual structure, doubling feature resolution through transposed convolution. The intermediate layer comprises two cascaded convolution layers, avoiding the insufficient non-linear feature extraction capability of a single convolution.

1.3 Feature Decoder

The decoder reconstructs features based on received encoder features to obtain dense pixel predictions. The proposed network' s decoder includes upsampling modules, reconstruction layers, and a classification layer (see Figure 1). The classification layer consists of a single convolution layer and bilinear interpolation operation, functionally consistent with the upsampling module, doubling feature map resolution and completing pixel-level semantic classification. The convolution output channel number of the classification layer is set to the total number of semantic categories. To enhance gradient propagation, reconstruction layers adopt a residual structure composed of two sequentially connected convolution layers and a 1$×$1 convolution on the residual path.

Since each reconstruction layer receives stacked features from the same and lower scales, the first convolution and residual layer of the reconstruction layer reduce feature map channels to match the output channel number of the encoder layer at the same level, while the second convolution maintains feature map resolution and channel number. All convolution layers in the network are followed by batch normalization and ReLU layers. Overall, the decoder comprises five reconstruction units composed of upsampling modules and reconstruction layers, plus one classification layer. The multi-level shallow and deep feature reuse pathways effectively assist semantic prediction, while the progressive feature scale restoration ensures structural symmetry between decoder and encoder.

Given that DenseNet has variants with different numbers of convolutional layers –DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-161 (with feature

channel growth rates of 32 for the first three and 48 for the last)—and increasing parameter complexity, when adopting different variants, the feature output channels at each downsampling stage align with the corresponding variant, and the input feature channel numbers of the decoder' s reconstruction units adjust accordingly.

1.4 Loss Function

The loss function is closely related to network fitting direction and convergence speed. The semantic segmentation field typically employs cross-entropy for training:

$$L_{\mathrm{CE}} = -\sum_{c=1}^{M} y_c \log(p_c)$$

where M is the number of categories, and y_c and p_c represent the ground truth label vector and predicted probability map for category c, respectively. Considering that target distribution across scales cannot be perfectly balanced, cross-entropy loss cannot adequately balance this sample discrepancy. Therefore, this paper additionally introduces an improved DiceLoss [16] term to enhance network learning capability:

$$L_{\mathrm{dl}} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i}$$

where p_i and g_i denote the binary predicted value and binary ground truth label value for the i-th pixel in the image' s pixel domain N, respectively. Thus, the total network loss is:

$$L_{\mathrm{total}} = \frac{1}{2}(L_{\mathrm{CE}} + L_{\mathrm{dl}})$$

Since the value ranges of the two loss terms share the same order of magnitude, they each occupy half the weight. These two terms jointly guide network learning, compensating for the shortcomings of using a single cross-entropy loss term.

2 Experiments

2.1 Dataset and Training Details

MFNet released the first pixel-level semantically annotated RGB-T urban road scene image dataset, containing 820 daytime and 749 nighttime RGB-infrared image pairs, all uniformly resized to 480$\times$640. The dataset manually annotates nine semantic classes on driving roads: Car, Person, Bike, Curve, Car Stop,

Guardrail, Color Cone, Bump, and Unlabelled background. Pixel counts across categories are extremely imbalanced, particularly for Car Stop and Guardrail. This paper follows the original dataset split, with a 2:1 ratio for training and validation sets (day/night images equally split), and the remaining 393 image pairs used as the test set.

The network model is implemented in PyTorch, using Stochastic Gradient Descent (SGD) as the optimizer. Network layers are initialized via the Xavier scheme [17], with the learning rate starting at 1×10^{-2} and decaying exponentially per epoch with a decay weight of 0.95. Input images are normalized to [0,1] and randomly flipped before each epoch to prevent overfitting. BatchSize is adjusted according to the backbone variant: 2 for DenseNet-161, 4 for DenseNet-201 and DenseNet-169, and 6 for DenseNet-121. All training and testing are conducted on a computer equipped with an NVIDIA GeForce RTX 3090 GPU (24GB VRAM), 32GB RAM, and an AMD Ryzen 9 5900X CPU. Training continues until the loss function ceases to decrease, with the best weights selected via the validation set. No processing is applied to inputs during testing.

2.2 Performance Evaluation Metrics

Segmentation performance is evaluated through both qualitative and quantitative means—visual comparison of segmentation results and numerical analysis via mean Accuracy (mAcc) and mean Intersection over Union (mIoU). mAcc measures the average probability of correctly classifying target image pixels across all semantic categories:

$$\mathrm{mAcc} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}$$

where N is the total number of categories (here $N = 9$), TP_i denotes the number of pixels correctly predicted as class i (true positives), and FN_i denotes pixels incorrectly predicted as non-i (false negatives). mIoU measures the average overlap between predicted segmentation and ground truth labels across all categories:

$$\mathrm{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i}$$

where FP_i represents pixels incorrectly predicted as class i (false positives). Both metrics positively correlate with segmentation performance.

2.3 Experimental Results Analysis

The proposed network' s segmentation performance is validated on the MFNet test set, with comparative methods involving current state-of-the-art RGB-T

segmentation models. All data are sourced from corresponding papers and open-source code. Table 1 and Figure 4 provide quantitative comparison results and visualization comparisons of day/night image sequences, respectively.

As shown in Table 1, the proposed segmentation network achieves the best values for both mAcc and mIoU metrics. Specifically, Car and Person semantic categories attain the highest overall metrics, likely benefiting from the combined effects of the nested architecture and deep difference feature fusion strategy. The former enhances learning capability for large-scale and easily identifiable objects, while the latter strengthens deep semantic expression for targets with significant feature differences—most notably benefiting Cars and Persons, which exhibit the greatest feature modality differences at night. For Curve, its white reflective properties provide slightly superior imaging advantages over thermal radiation information at night, somewhat enhancing its inherent feature strengths. In contrast, the Bike category, often densely clustered in multiple scenes, may suffer from overfitting in dense intermediate layers, weakening segmentation advantages for individual bikes and preventing the best accuracy. Small-scale objects like Color Cones likely benefit from this approach, as evidenced by MFNet and RTFNet, whose models lack information flow channels bridging encoders and decoders, resulting in insufficient feature learning capability for small objects. AFNet and MLFNet, through co-attention fusion and multi-level encoder feature skipping advantages respectively, both demonstrate excellent segmentation capabilities across scales. For other categories, Guardrail and Car Stop have insufficient sample counts in the test set (Guardrail appears in only 4 of 393 image pairs), causing poor segmentation performance across all models—particularly MFNet and RTFNet—likely due to loss of already scarce feature information during training in networks lacking feature reuse or adjustment mechanisms.

Further detailed differences can be observed in Figure 4. Taking columns 2 and 8 as examples, the Bike category shows segmentation results closest to ground truth.

To further investigate model segmentation efficacy, Table 2 lists experimental comparison results on daytime and nighttime images separately.

As shown in Table 2, all methods achieve better segmentation performance at night. This may be because under well-lit conditions, RGB images already contain rich detail information for easy segmentation, and incorporating thermal radiation information can cause hedging effects on some dominant features, weakening their semantic expression. At night, the two modalities exhibit a larger semantic gap, making infrared information incorporation more effective for improving semantic partitioning results.

Comparing day/night test sequences, the proposed method demonstrates better mean accuracy and mIoU in nighttime scenes, indirectly corroborating that the proposed deep difference feature fusion strategy can fully integrate RGB and infrared features, as infrared images naturally possess imaging advantages at night when feature differences become more pronounced.

2.3.1 Encoder Backbone Variants Different DenseNet variants as encoder backbones yield varying segmentation performance. To investigate this impact, we retrain the network with only the backbone variant changed until loss ceases to decrease. Figure 5 shows performance on the MFNet test set.

In Figure 5, mFPS denotes mean frames per second on the test set. To align with the increasing direction of segmentation metrics, the actual inverse of mFPS is plotted. Results show that as DenseNet variant complexity increases, the proposed network exhibits increasing trends in both accuracy and IoU metrics. In contrast, the average time consumed per segmented frame correlates approximately only with network depth. This suggests that deeper architectures with increased parameters possess stronger segmentation learning capability, while network inference speed is primarily affected by network depth.

2.3.2 Encoding Feature Reuse Methods In the proposed model, encoders and decoders are connected via nested upsampling and intermediate layers, enabling dense reuse of shallow and deep encoder features. To validate this approach, we remove all information reuse paths between encoding and decoding components, retaining only the connection between the encoder' s final layer and decoder (denoted as U-shaped Direct Connection). Additionally, we add skip connections to the U-shaped structure to transmit same-level encoder features to corresponding decoder reconstruction layers (denoted as Same-level Skip Connection). Using DenseNet-161 as the backbone and keeping other conditions unchanged, we retrain until convergence. Table 3 presents segmentation comparison results on the MFNet test set.

As shown in Table 3, network segmentation performance degrades sharply when decoder features are not reused. Performance improves when same-scale features are reused via long connections, particularly in accuracy. When multi-scale shallow and deep features are reused through nested connections, accurate segmentation coverage further improves, though per-pixel segmentation precision slightly degrades. In summary, reusing encoder features substantially impacts segmentation performance. Dense reuse of shallow and deep features most effectively improves mIoU but slightly weakens accuracy improvement, possibly because dense intermediate connections cause segmentation overfitting in some scenes.

2.3.3 Deep Feature Fusion Strategy To validate the effectiveness of the deep difference feature fusion strategy, we compare two alternative fusion strategies: the self-similarity fusion unit from Transformers [19] and pixel-difference-based complementary fusion. The former focuses on spatial correlations between pixel positions within feature maps themselves, resembling a position-attention mechanism, while the latter focuses on semantic correlations between multi-modal features at the pixel level. In contrast, our proposed strategy focuses on semantic correlations between vector features of RGB and infrared feature maps. Table 4 presents ablation experimental results for these three fusion strategies

on the MFNet test set.

As shown in Table 4, the proposed fusion strategy provides optimal fusion guidance for RGB and infrared deep features. This is because, in multi-modal feature fusion, self-similarity fusion strategies neglect expression of dissimilar image features, while pixel-difference-based fusion only focuses on local feature correlations—both having limitations in integrating high-dimensional features of multi-modal images with property differences. In summary, for high-dimensional abstract feature fusion of multi-modal objects with different imaging mechanisms, mining their respective distinct features and performing targeted feature-level complementary fusion yields fused features with more robust semantic expression.

3 Conclusion

This paper designs a nested semantic segmentation network that fuses deep difference features from RGB and infrared images. Considering that features from different encoding scales possess semantic representations at various levels, the model constructs nested intermediate pathways to achieve efficient dense reuse of shallow and deep features. Simultaneously, to enhance the semantic expressiveness of high-dimensional abstract features from RGB and infrared images, a deep difference feature fusion strategy is designed to achieve feature complementary enhancement. Comparative experiments with state-of-the-art models on public datasets demonstrate the proposed model' s superior segmentation performance, with ablation experiments validating the effectiveness of dense feature reuse and deep difference feature fusion strategies.

Future work will focus on optimizing the combination of difference feature fusion strategies and attention mechanisms to improve segmentation accuracy for complex objects. Additionally, we consider generalizing the RGB-T segmentation network to other multi-modal image semantic segmentation domains.

References

[7] Ha Q, Watanabe K, Karasawa T, et al. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes [C]// 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017.

[8] Hazirbas C, Ma L, Domokos C, et al. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture [C]// Asian conference on computer vision. Springer, Cham, 2016: 213-228.

[9] Sun Y, Zuo W, Liu M. RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes [J]. IEEE Robotics and Automation Letters, 2019: 2576-2583.

[10] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition

[J]. IEEE, 2016.

[11] Xu J, Lu K, H Wang. Attention Fusion Network for Multi-spectral Semantic Segmentation [J]. Pattern Recognition Letters, 2021 (4).

[12] Guo, Z., Li, X., Xu, Q., Sun, Z.: Robust semantic segmentation based on rgb-thermal in variable lighting scenes. Measurement 186, 110176 (2021).

[13] Zhou Z, Siddiquee M, Tajbakhsh N, et al. UNet+: A Nested U-Net Architecture for Medical Image Segmentation [J]. 2018.

[14] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation [C]// International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[15] Huang G, Liu Z, Laurens V, et al. Densely Connected Convolutional Networks [J]. IEEE Computer Society, 2016.

[16] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation [C]// 2016 fourth international conference on 3D vision (3DV). IEEE, 2016: 565-571.

[17] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [C]// Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010: 249-256.

[18] Lan, X., Gu, X. & Gu, X. MMNet: Multi-modal multi-stage network for RGB-T image semantic segmentation. Appl Intell (2021). https://doi.org/10.1007/s10489-021-02687-7.

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need [J]. arXiv, 2017.

Note: Figure translations are in progress. See original paper for figures.