From Contexts to Locality: Ultra-high Resolution Image Segmentation via Locality-aware Contextual Correlation

Qi Li[1],Weixiang Yang[1],Wenxi Liu[1*],Yuanlong Yu[1*],Shengfeng He[2]
[1]*College of Mathematics and Computer Science, Fuzhou University**
[2]*School of Computer Science and Engineering, South China University of Technology*

Abstract

Ultra-high resolution image segmentation has raised increasing interests in recent years due to its realistic applications. In this paper, we innovate the widely used high-resolution image segmentation pipeline, in which an ultra-high resolution image is partitioned into regular patches for local segmentation and then the local results are merged into a high-resolution semantic mask. In particular, we introduce a novel locality-aware contextual correlation based segmentation model to process local patches, where the relevance between local patch and its various contexts are jointly and complementarily utilized to handle the semantic regions with large variations. Additionally, we present a contextual semantics refinement network that associates the local segmentation result with its contextual semantics, and thus is endowed with the ability of reducing boundary artifacts and refining mask contours during the generation of final high-resolution mask. Furthermore, in comprehensive experiments, we demonstrate that our model outperforms other state-of-the-art methods in public benchmarks. Our released codes are available at https://github.com/liqiokkk/FCtL.

1. Introduction

With the advance of photography and sensor technologies, the accessibility to ultra-high resolution images (i.e., 2K, 4K, or even higher resolution images) has opened new horizons to the computer vision community. It will benefits a wide range of imaging applications, e.g., urban planning and remote sensing based on high-resolution geospatial images and high-resolution medical image analysis, and thus the demand for studying and analyzing such images has urgently increased in recent years.

In this paper, we aim at the specific task of semantic segmentation for ultra-high resolution geospatial images captured from aerial view. The recent development of deep convolutional neural networks (CNNs) has given rise to remarkable progress of semantic segmentation techniques.

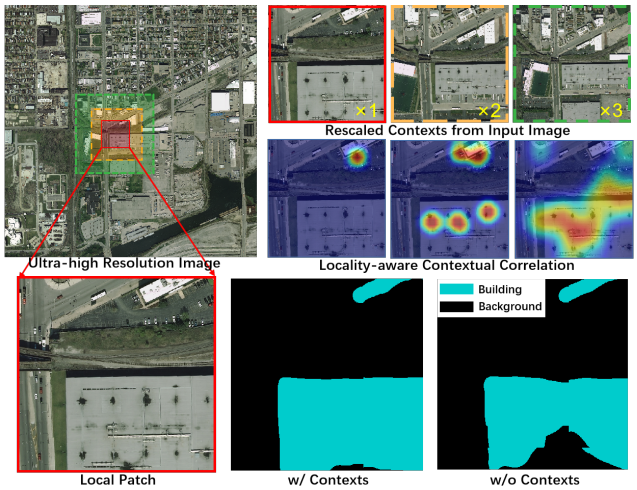


Figure 1. For the task of ultra-high resolution image segmentation, the most common approach is to segment cropped local patches and then combine them into a high-resolution mask. To address the core problem on local segmentation quality, we propose a locality-aware contextual correlation based model that exploits rescaled various contexts ($\times 1$, $\times 2$, $\times 3$ large as local patch in the original image) to produce refined results.

Yet, most CNN-based segmentation models target on full resolution images and perform pixel-level class prediction, which requires more computation resources comparing to image classification and object detection. This hurdle becomes significant when the image resolution grows to be ultra high, leading to the pressing dilemma between memory efficiency (even feasibility) and segmentation quality.

Particularly, in order to segment an ultra-high resolution image, the prevailing practice is either to downsample it to a smaller spatial dimension before performing segmentation, or to separately segment the partitioned patches and merge their results into a high-resolution one. These trivial practices sacrifice the segmentation quality for the model efficiency. Additionally, the recent attempts propose to utilize the well-pretrained segmentation models to obtain the coarse segmentation masks and another model to refine the contours of the masks [5, 37]. However, these methods mainly focus on high-resolution natural images or daily photos concerning with large objects, while the high-

*Wenxi Liu and Yuanlong Yu are the corresponding authors.

resolution geospatial images are captured from aerial views covering a large field of view, which may contain many objects/regions with large contrast in scale and shape. Hence, it requires the segmentation model to be capable of capturing not only the semantics over large image regions but also the image details of different granularity. The recent work GLNet [4] proposes to incorporate the local and global information via a two-stream network that separately processes the downsampled global image and cropped local patches, as well as a feature sharing module that shares the concatenated local and global features in both streams. Their method can achieve obvious improvements over existing methods, which embodies the importance of contextual information for segmentation performance. Nevertheless, their feature sharing scheme does not spatially associate local features with the global ones and thus does not well exploit their correlation, which makes their model too complex to optimize and their performance suboptimal.

To thoroughly utilize the rich information within ultra-high resolution geospatial images, we present an ultra-high resolution geospatial image segmentation model featuring with the locality-aware contextual correlation scheme. Similar to [4, 25], our framework is based on the widely used practice for high-resolution image segmentation, in which image patches are regularly cropped from the original image, then individually segmented, and finally their local results are overlayingly merged. However, each local patch of the ultra-high resolution geospatial images often contain semantic regions with large contrast in sizes (e.g., house and forest), which challenges the local segmentation model. Inspired by prior practices (e.g. [4]), contextual information turns out effective to resolve this problem. But, unlike previous methods, we propose that the semantics within local patches can be structually and complementarily associated and inferred by their contextual regions of different scales. For instance, in Fig. 1, the contexts with varied coverage guide the model to the attentive regions relevant to the objects of different granularity in the image (e.g., small or large building). Hence, we propose a locality-aware contextual correlation based deep network model to exploit the correlation between local patch and its contextual regions. In concrete, we first present a locality-aware contextual correlation module to capture the positional relevance of local patch and context, which is enabling to attentively enhance the relevant features of local patch, i.e., *locality-aware features*. Then, we propose an adaptive context fusion scheme to balance and combine the locality-aware features associated by various contexts. As shown in Fig. 1, the contexts can lead to different yet complementary locality-aware features, thus allow tolerance to misleading information in a single context. To do so, the corresponding spatial weight maps of different locality-aware features are predicted on-the-fly to accomplish the complementary fusion.

Furthermore, to obtain the final segmentation result of the ultra-high resolution image, the results of local patches will be put back together. Directly montaging local segmentation masks may cause boundary vanishing artifacts for adjacent patches, so the prior practice is to overlap adjacent patches partially and compute the average results for overlap regions. To some extent, this trivial approach can reduce the artifacts, yet cannot achieve the optimal results. Therefore, we propose an effective contextual semantics refinement network that utilizes the correlation of local mask and context mask to enhance the relevant semantic regions and thus adaptively refine the local results without introducing boundary vanishing artifacts. Besides, our proposed model can also leverage the contextual semantics to polish the contours of segmentation masks.

To evaluate our model, we conduct comprehensive experiments and demonstrate that our proposed model outperforms the state-of-the-art approaches on public ultra-high resolution arerial image datasets, DeepGlobe and Inria Aerial. The main contributions of our paper are summarized as below:

- We present an ultra-high resolution image segmentation framework based on a novel local segmentation model. It leverages the locality-aware contextual correlation and the adaptive feature fusion scheme, which associates and combines local-context information to strengthen local segmentation.
- We present a contextual semantics refinement network that leverages the relevance of local segmentation and context mask to avoid boundary vanishing artifacts and refine the local semantic mask.
- Our method achieves the state-of-the-art semantic segmentation performance in several public ultra-high resolution geospatial image datasets.

2. Related Works

Semantic Segmentation. In recent years, semantic segmentation has achieves remarkable progress [2,7,10–12,18, 19,27,30]. Fully convolutional network (FCN) [18] was the first CNN architecture adopted for high-quality segmentation. U-Net [25] used skip-connections to concatenate low level features to high-level ones. Similar structures were also adopted by [1, 22]. Unfortunately, these models suffer from prohibitively high GPU memory demand for ultra-high resolution images. ENet [23] and ICNet [34] reduced GPU memory via model compression. However, these models were not effective on ultra-high resolution images. Recently, CascadePSP [5] is proposed to refine the coarse segmentation results from a pretrained model to generate high-quality results. GLNet [4] preserves both global and local information and interact each other through deeply

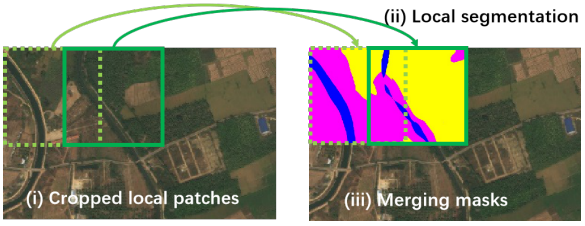


Figure 2. Main procedure of high-resolution image segmentation, consisting of (i) cropping local patches from the ultra-high resolution image; (ii) local patch segmentation; (iii) merging local masks into a high-resolution mask.

shared layers, which is able to balance its performance and GPU memory usage. Compared with GLNet, the key difference rests in our proposed multi-context based local segmentation model, while GLNet relies on the holistic image as the only context and simply concatenates the local and cropped global features for segmentation. Additionally, we also present a new contextual refinement model for merging local results into a HD one, which has not been studied.

Multi-Scale and Context Aggregation. Multi-scale information [2, 3, 9, 16, 31, 36, 38] has proven to be effective for segmentation, via integrating high-level and low-level features to capture patterns of different granularity. RefineNet [13] introduced a multi-path refinement block to combine multi-scale features via upsampling lower-resolution features. [8] adopted a Laplacian pyramid to utilize higher-level features to refine boundaries reconstructed from lower-resolution maps. Feature Pyramid Networks (FPN) [14] progressively upsampled feature maps of different scales and aggregated them in a top-down fashion. On the other hand, context aggregation also plays a key role in encoding the local spatial neighborhood, or even non-local information [2, 4, 17, 28, 29, 32, 35]. ParseNet [17] incorporated global pooling to aggregate different levels of contexts. DeepLab [2] proposed dilated convolution and atrous spatial pyramid pooling module to aggregate global contexts into local information. In recent works [4, 21, 24, 33], the deep/shallow branches are combined to aggregate global context and high-resolution details. Unlike previous works, we propose that the local segmentation can be spatially correlated with various contexts, and we propose an adaptive fusion scheme to combine different locality-aware features.

3. Methodology

Our proposed ultra-high resolution image segmentation framework follows the three-step procedure as shown in Fig. 2), which is consistent with the common practice applied in prior works (e.g., [4, 25]). First, given an ultra-high resolution image $\dot{I}$ with width W and height H, we evenly partition it into N local patches $\{I_k\}$ ($k = [1, \cdots, N]$, $I_k \subset \dot{I}$) with width w and height h ($w < W$ and $h < H$). Next, a local semantic segmentation model computes the local result for each patch. Last, we merge the local results into one piece as the final high-resolution segmentation mask. Our main contributions rest in how to generate fine local segmentation (the second step) and refined results that can be seamlessly merged into a high-resolution mask (the third step). As follows, we will elaborate the technical details.

3.1. Our Proposed Local Segmentation Model

As the core of our ultra-high resolution segmentation framework, we propose a novel local segmentation model to process each cropped patch (Fig. 3). Yet, each local patch only covers a confined field of the ultra-high resolution image, which often contains regions of varied scales or truncated objects, and thus it tends to deliver incomplete information and may easily cause erroneous semantic segmentation. To address this concern, we propose a locality-aware contextual correlation based segmentation model for processing each local patch.

As illustrated in Fig. 3, our local segmentation model is based on a multi-stream encoder-decoder architecture, consisting of the feature extraction modules (i.e., encoder), locality-aware contextual correlation module, multi-context fusion module, and decoder. In specific, a local patch along with contexts of different scales, which are rescaled into the same size for reducing computation overhead, are fed into the network for feature extraction. Then, the features of contexts are separately associated with the features of local patch via the locality-aware contextual correlation module and adaptively fused. In final, the features will be upsampled to obtain the local segmentation mask.

As follows, we will first introduce how to choose the context of local patch, and then describe the locality-aware contextual correlation module and multi-context fusion scheme.

3.1.1 Context of Local Patch

Regarding of the k-th patch I_k, U_k is denoted as another image region within the input image $\dot{I}$, which is not smaller than and covers I_k. U_k has the width w_u and height h_u, s.t. $w \leq w_u \leq W$ and $h \leq h_u \leq H$.

Given a local patch, there are many candidate context regions. In practice, we design the following three types of context regions. (1) We set the size of the candidate context subject to $w_u = \lambda w$ and $h_u = \lambda h, (\lambda \geq 1, w_u \leq W, h_u \leq H)$, and its center aligned with the center of the local patch (see the examples in Figs. 1 and 3). (2) The largest context we can utilize is exactly the whole image, i.e., $U_k \equiv \dot{I}$, dubbed *global context*. (3) The smallest context is the patch itself, dubbed *local context*, i.e., $w_u \equiv w$ and $h_u \equiv h$. Generally, the larger contexts offer more contextual cues that

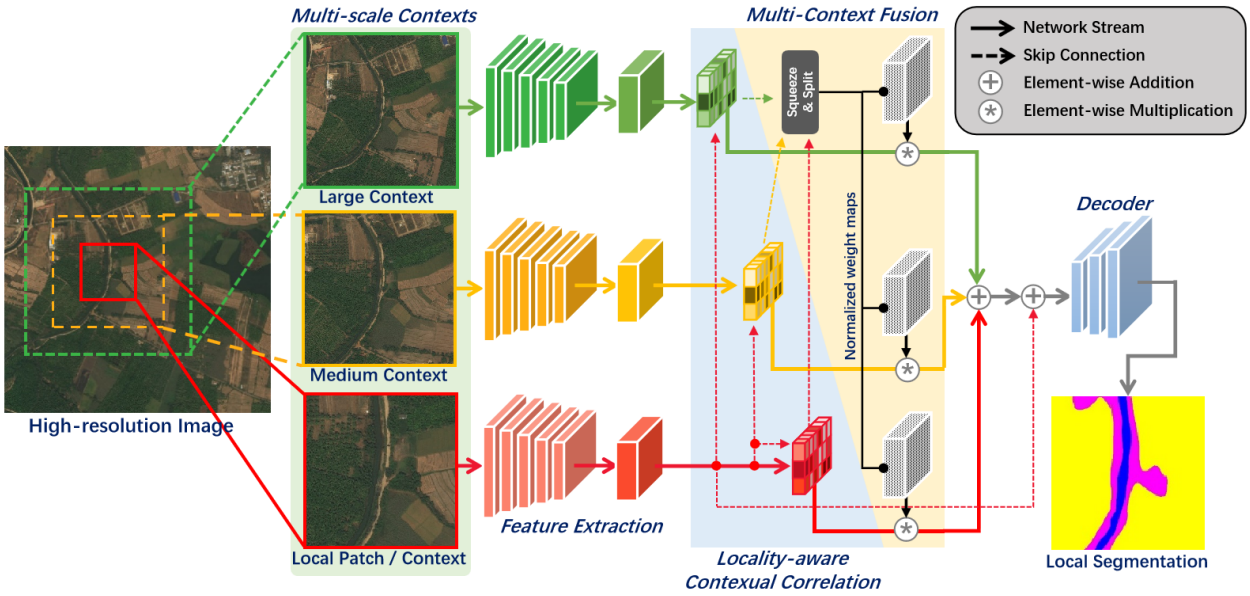


Figure 3. Illustration of our local segmentation model. In specific, a certain local patch cropped from the high-resolution image and its contexts are passed into the network branches separately to extract features and then measure their relevances against local patch to obtain the locality-aware features. Last, these features are adaptively fused for producing high-quality local segmentation results.

may be responsible for large regions or objects, while the smaller ones provide more details that may be attributed to small regions or objects. Before feeding into network, the contexts will be normalized as the same dimension as the local patch.

3.1.2 Locality-aware Contextual Correlation

To strengthen the segmentation of a local patch, we would like to associate the contextual information with the local information. Thus, we propose a locality-aware contextual correlation module $\mathcal{F}_{lcc}$ to evaluate the relevance between the features of I_k and U_k, and then utilize it to obtain its locality-aware features.

The structure of our produced module is specified in Fig. 3. First, the features of I_k and U_k are separately extracted through the same network structure (i.e., Conv1 to Conv3 of the pretrained VGG16 [26]), denoted as $\mathbf{X}_k^i$ and $\mathbf{X}_k^u$ ($\mathbf{X}_k^i, \mathbf{X}_k^u \in \mathbb{R}^{c \times h_x \times w_x}$). Next, the relevance of I_k and U_k is calculated via the inner product of local features $\mathbf{X}_k^i$ and contextual features $\mathbf{X}_k^u$, i.e., $\mathbf{R}_k = \langle \mathbf{X}_k^i, \mathbf{X}_k^u \rangle$, which measures the non-local correlation by establishing the pairwise pixel-level relation for $\mathbf{X}_k^i$ and $\mathbf{X}_k^u$. Hence, the relevance can be further applied as an attention map to enhance the local features, $\mathbf{X}_k^i$, which attends to the semantic regions more relevant with the locality. Specifically, $\mathbf{R}_k$ is passed through a softmax layer to obtain the attention map and then perform inner product with $\mathbf{X}_k^i$, i.e., $\overline{\mathbf{X}}_k = \langle \text{Softmax}(\mathbf{R}_k), \mathbf{X}_k^i \rangle$. For the sake of clarity, we denote the procedure as $\overline{\mathbf{X}}_k = \mathcal{F}_{lcc}(I_k, U_k)$.

3.1.3 Multi-context Fusion Module

For the ultra-high resolution geospatial images that often contain a large number of objects with large size variations, the contexts of different scales may be attributed to the segmentation of the objects with various granularity. Therefore, properly combining different contextual information can be complementary for extracting semantics and removing artifacts.

Specifically, we assume that there are T corresponding context regions that may affect the local segmentation. Formally, given the patch I_k, we have several corresponding context regions, U_k^t ($t = [1, \cdots, T]$) and pass them into each stream of our local segmentation model to obtain locality-ware features $\overline{\mathbf{X}}_k^t$ ($\overline{\mathbf{X}}_k^t = \mathcal{F}_{lcc}(I_k, U_k^t)$). To effectively exploit and combine the locality-aware features that stem from various contexts, as shown in Fig. 3, we integrate a multi-context fusion scheme into our local segmentation model. Here, we propose a novel network module $\mathcal{F}_{est}$ that estimates the weight maps $\{\mathbf{H}^t\}$ corresponded to the features $\{\overline{\mathbf{X}}_k^t\}$, respectively.

Specifically, the locality-aware features are first concatenated and passed through a *squeeze-and-split* structure, which compresses and entangles the multi-scale features before predicting the normalized weights for the features from different sources, i.e., $\{\mathbf{H}^t\} = \mathcal{F}_{est}(\{\overline{\mathbf{X}}_k^t\})$. In particular, the squeeze-and-split structure squeezes the concated features via a convolutional layer with the kernel size 1×1, which blends the locality-aware features. Then, the squeezed features are reconstructed to the original dimension via another 1×1 convolutional layer and passed through a softmax to obtain T normalized weight maps

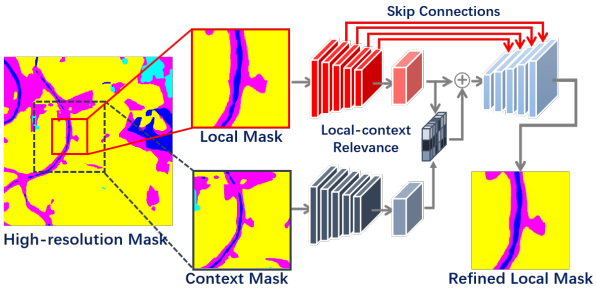


Figure 4. Illustration of contextual semantics refinement network. Given a crude high-resolution semantic mask, we feed a local mask and its context mask into a two-branch network to refine the local mask.

$\{\mathbf{H}^t\}$ ($\mathbf{H}^t \in \mathbb{R}^{h_x \times w_x}$, $t = [1, \cdots, T]$) which balance the contributions of each term. Hence, we have the fused features $\overline{\overline{\mathbf{X}}}_k$ as below:

$$\overline{\overline{\mathbf{X}}}_k = \sum_{t=1}^{T} \mathbf{H}^t \odot \overline{\mathbf{X}}_k^t,$$
$$s.t. \sum_{t=1}^{T} \mathbf{H}^t = \mathbf{1}, \quad (1)$$

where $\odot$ refers to the element-wise multiplication and $\mathbf{1}$ represents the matrix where all elements are 1. Note that, the elements of T weight maps along each channel are summed to be 1.

Last, the fused features will be joined with the features of the local patch via a skip connection to form a residual structure, and then used to compute the segmentation mask of the patch via several upsampling layers in the decoder. In this way, our feature fusion scheme is able to take advantage of the complementary information from different contexts. Last, we apply the focal loss [15] as the objective function, in which γ is set as 3.

3.2. Contextual Semantics Refinement Network

In prior practices, montaging the computed segmentation masks of all cropped patches easily cause boundary artifacts, so the local masks are stacked together in an overlapping manner (see the third step in Fig. 2) and the results on the overlap regions are computed in average, which somewhat reduces the artifacts. Merging all the local masks in this way eventually leads to the final high-resolution result. However, this trivial approach can hardly achieve the optimal results without adaptively considering the semantic correlation between local patch and its context. To address this issue, we propose a contextual semantics refinement network to exploit contextual semantic mask to refine the local mask.

Given the computed local segmentation masks, we can spawn the crude high-resolution mask via simply montaging local patches or merging masks in an overlapping manner. Although this coarse result may contain artifacts, the semantics from context play the important role to embody the geospatial layout of neighboring regions, and thus facilitate the refinement of the local mask. In specific, as shown in Fig. 4, our refinement network bases on a two-stream variant of U-Net architecture which incorporates a local-context relevance module to associate the context and local mask. The local-context relevance module is similar to the locality-aware correlation module in our local segmentation model, which measures the correlation between the features of local mask and context mask that are rescaled to be the same dimension as local mask, and then attentively enhances the local mask. To this end, the contextual semantics can be leveraged to not only remove the boundary vanishing artifacts but also improve the contours of the local mask. Besides, same as the standard U-Net structure, our network bridges the features of the downsampling and upsampling layers via skip connections that deliver low level details to deep layers so as to achieve the refined results. Moreover, we adopt the focal loss as the objective function as well. All the refined local results can be further applied to montage a better quality high resolution semantic mask.

4. Experimental Results

In this section, we demonstrate the comprehensive experimental results over public benchmarks. We thoroughly compare our model against the state-of-the-art methods to show the segmentation quality and conduct the ablation study to evaluate the capability of our model.

4.1. Datasets

DeepGlobe [6]. This dataset contains 803 ultra-high resolution images (2448×2448 pixels). Following [4], we split images into training, validation and testing sets with 455, 207, and 142 images respectively. The dense annotation contains seven classes of landscape regions, including cyan represents "urban", yellow represents "agriculture", purple represents "rangeland", green represents "forest", blue represents "water", white represents "barren", where one class out of seven called "unknown" region is not considered in the challenge.

Inria Aerial [20]. This dataset covers diverse urban landscapes, ranging from dense metropolitan districts to alpine resorts. It provides 180 images (from five cities) of 5000×5000 pixels, each annotated with a binary mask for building/non-building areas. Unlike DeepGlobe, it splits the training/test sets by city. We follow the protocol as [4] by splitting images into training, validation and testing sets with 126, 27, and 27 images, respectively.

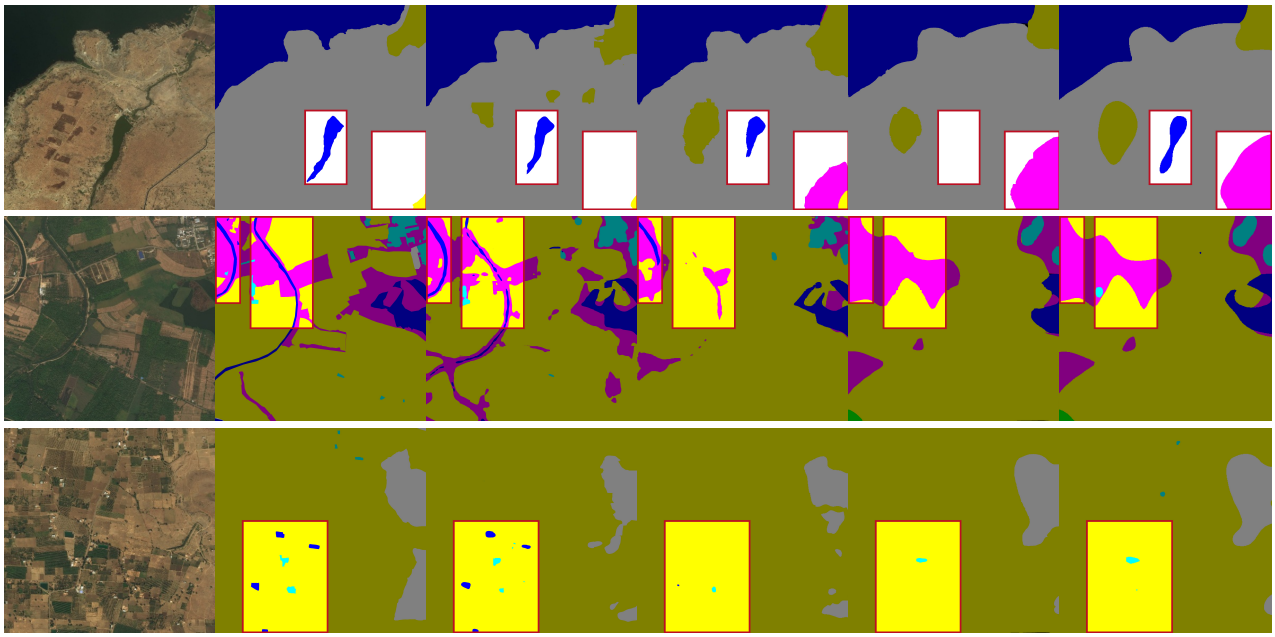


Figure 5. We illustrate several examples of semantic segmentation in ultra-high resolution images, comparing with the state-of-the-arts. In the figures, masks with varied colors represent different semantic regions. Particularly, cyan represents "urban", yellow represents "agriculture", purple represents "rangeland", green represents "forest", blue represents "water", and white represents "barren".

4.2. Implementation Details

Settings for contexts. In practice, we apply three contexts in our model, which are denoted as local, medium, and large contexts. The sizes of contexts differ for two benchmarks. We evaluate the performance of different context settings in Sec. 4.4.

Training details. We implement our framework using Pytorch on a computer with a single NVIDIA GTX 1080Ti GPU. In particular, we adopt VGG16 [26] as our backbone and our baseline model is similar to FCN-8s [18]. All the input images (i.e., local patches) are normalized to 508×508 and the output size is 508×508, which follows the setting of [4] in order to trade-off performance and efficiency. When merging local results into a high-resolution one, we let neighboring patches have a 120×508 overlapping region to avoid boundary vanishing.

During training our local segmentation model, we adopt the Adam optimizer and a mini-batch size of 6 by gradient accumulation. The initial learning rate is set to 5×10^{-5} and it is decayed by a poly learning rate policy where the initial learning rate is multiplied by $(1 - \frac{iter}{total\_iter})^{0.9}$ after each iteration. In practice, it takes 50 epochs to converge our model. Besides, as our baseline and comparison method, FCN-8s [18] also follows the training strategy above.

Regarding the independent training of the contextual semantics refinement network, we adopt similar training settings as the local segmentation model. Note that, the inputs of the refinement network (i.e., local and context masks) stem from our trained segmentation model. Once the refinement network shares the same training dataset as local segmentation model, it can easily lead to the overfitting of refinement network and degrades its inference performance. To address this problem, we apply a simple strategy. Instead of thoroughly training the local segmentation model, we early stop its training after 20 epochs and leverage the non-converged segmentation model to generate samples to train the refinement network, which turns out to work well.

4.3. Comparison with State-of-the-arts

For evaluation, we compare our approach against U-Net [25], ICNet [34], PSPNet [35], SegNet [1], DeepLab v3+ [2], FCN-8s [18], CascadePSP [5], and GLNet [4] over the benchmarks DeepGlobe and Inria Aerial, in terms of mIOU(%), F1(%), and Accuracy(%). Their results are depicted in Table 1 and Table 2, in which we follow most of the quantitative results provided by [4]. Most of these methods are not designed for ultra-high resolution images (denoted as *Generic Model* in Table 1), so there are two ways to train these models: 1) training from the local patches and merging local results; and 2) training from the downscaled global images. Thus, we provide their corresponding metrics on DeepGlobe as *Local Inference* and *Global Inference* in Table 1. CascadePSP and GLNet are specifically tailored for the task of ultra-high resolution image semantic segmentation (denoted as *High-Res Model* in Table 1). In particular, CascadePSP mainly aims to handle natural images, while GLNet can be applied for geospatial images. Note that, in Table 1, GLNet* refers to the model without its global-local feature sharing module. All of the results are obtained following the same training and testing protocol. Besides, since the original paper of CascadePSP has

Generic Model	*Local Inference*			*Global Inference*		
	mIOU	F1	Acc.	mIOU	F1	Acc.
U-Net [25]	37.3	-	-	38.4	-	-
ICNet [34]	35.5	-	-	40.2	-	-
PSPNet [35]	53.3	-	-	56.6	-	-
SegNet [1]	60.8	-	-	61.2	-	-
DeepLab v3+ [2]	63.1	-	-	63.5	-	-
FCN-8s [18]	71.8	82.6	87.6	68.8	79.8	86.2
GLNet* [4]	57.3	64.6	72.2	66.4	79.5	85.8
High-Res Model	mIOU		F1		Acc.	
CascadePSP [5]	68.5		79.7		85.6	
GLNet [4]	71.6		83.2		88.0	
Ours	**73.5**		**83.8**		**88.3**	

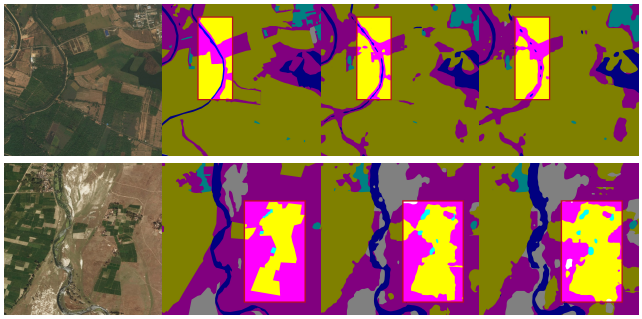
Table 1. Comparison with state-of-the-arts on DeepGlobe.

Model	mIOU	F1	Acc.
ICNet [34]	31.1	-	-
DeepLab v3+ [2]	55.9	-	-
FCN-8s [18]	69.1	81.7	93.6
CascadePSP [5]	69.4	81.8	93.2
GLNet [4]	71.2	-	-
Ours	**73.7**	**84.1**	**94.6**

Table 2. Comparison with state-of-the-arts on Inria Aerial.

Context			mIOU	
Local	Medium	Large	DeepGlobe	Inria Aerial
			71.84	69.08
✓			72.12	72.50
	✓		72.67	72.48
		✓	72.67	72.46
	✓	✓	73.12	73.18
✓	✓	✓	**73.22**	**73.53**

Table 3. Efficacy of contexts for local segmentation.



Input Image Ground-truth w/ Contexts w/o Contexts

Figure 6. Examples show the efficacy of contexts.

not reported the results on these two datasets, we train their model following the same protocol in our experiments as well. CascadePSP requires a pretrained model to provide rough global results. In our experiments, the performances of their pretrained global models are 66.9% and 69.0% in DeepGlobe and Inria Aerial, respectively. As observed in Table 1 and 2, amongst all comparison methods, our model achieves the state-of-the-art performance comparing to the competing methods in the respective datasets. Our task often suffers from severe class imbalance problem, e.g., the category "agriculture" occupies much more area (i.e. pixels) than "water". The pixel-wise metrics F1 and accuracy can hardly reflect how the models handle this problem. Instead, mIOU measures the average segmentation quality of each category. Thus, the major improvements on mIOU (more than $\sim 2\%$) indicates the effectiveness of our model. In addition, we show several qualitative comparison results in Fig. 5. As observed, our model is able to identify strip-shaped regions (e.g., river) and small regions (e.g. agriculture), which is benefited from the correlation between local and contexts.

4.4. Ablation Study

In this section, we delve into the modules and settings of our proposed model and demonstrate their effectiveness.

Efficacy of contexts. Our baseline is based on FCN-8s [18] without introducing contexts. As shown in Table 3, in general, integrating contexts obviously improves the segmentation performance, which boosts the performance from 71.84% to 73.22% in DeepGlobe and from 69.08% to 73.53% in Inria Aerial. Particularly, the smallest context (i.e., local context or local patch itself) provides the non-local self-correlation cues. Yet, the self-correlation features of local patch can hardly provide sufficient information to further infer the semantics of the patch. On the other hand, the medium and large contexts bring in the scaled contextual information and thus facilitate the segmentation. But, relying on medium context or large context alone may not always give rise to better results. For instance, for Inria Aerial, the results from the medium or large context are even slightly worse than the one from local context. Hence, exploiting the complementary information from the contexts of different scales can lead to better results. In Fig. 6, we show several examples produced by the models with or without contexts.

Locality-aware contextual correlation. For validation, we replace it with the naive local-global feature concatenation for comparison, in which our proposed scheme is superior to feature concatenation (73.5% vs 72.2% on DeepGlobe; 73.7% vs 72.8% on Inria Aerial in mIOU).

Scales of contexts. We investigate how context sizes affect the segmentation performance. We assess the models with different context sizes for both benchmarks in Table 4. Intuitively, if the context size is close to the size of local patch, the local patch and context are highly overlapped and they share too much redundant information that may not bring much performance gain. Therefore, for the sizes of candidate contexts, we choose the multiples of the local patch size (508×508) as the sizes of three contexts (i.e., small, medium, and large). For DeepGlobe, the optimal context sizes are 508×508, 1524×1524, and 2448×2448, in which the large context is exactly the entire image (i.e., the global context). For Inria Aerial, the optimal context sizes are 508×508, 1016×1016, and 1524×1524. The different configurations of these two datasets are due to the characteristics of their images. DeepGlobe includes the

	Context Size (pix.)			mIOU		Context Size (pix.)			mIOU
	Small	Medium	Large			Small	Medium	Large	
DeepGlobe	508	1016	2448	72.33	Inria Aerial	508	762	1524	72.85
	508	1524	2032	71.97		508	1016	1270	72.95
	508	**1524**	**2448**	**73.22**		**508**	**1016**	**1524**	**73.53**
	508	2032	2448	72.35		508	1270	1524	72.91
	1016	1524	2448	72.18		508	1016	2032	72.77
						762	1016	1524	72.85

Table 4. Scales of contexts for local segmentation.

geospatial images with different terrains (e.g., water and forest), while Inria Aerial contains top-down urban views, in which a large number of buildings can be observed. Therefore, for DeepGlobe, with the entire image as our large context can help better understand the semantics. On the contrary, for Inria Aerial, too large context after being rescaled into a smaller size will lose details of cities and makes the model hard to discern the buildings in the images, which thus causes the performance degradation.

Multi-context fusion. To show the advantage of our fusion scheme, we compare our module against two trivial fusion methods: 1) simply averaging the locality-aware features (i.e., $\frac{1}{T}\sum_{t=1}^{T}\overline{\mathbf{X}}^t$); and 2) offline estimating the optimal weights of locality-aware features (i.e., the estimated weights remain constant for each dataset). The comparison analysis results in the term of mIOU are demonstrated in Table 5. As observed, our adaptive fusion scheme achieves the best performance over the trivial fusion methods. In Fig. 7, we illustrate the exemplar results of our fusion scheme comparing to the results without fusion. Note that, our fusion scheme is essentially the generic version of the average fusion and weighted fusion, so its advantage shown on Table 5 may not be significant yet indicates the effectiveness.

Contextual semantics refinement network. We evaluate the effectiveness of the contextual semantics refinement network. We compare our approach to the trivial methods using non-overlapping montaging and overlapping merging (i.e. averaging). As shown in Table 6, our semantics refinement network surpasses the trivial methods in the term of mIOU. Besides, we study how the context size of this network influences the refinement network. In specific, we evaluate the results computed by the context sizes 762×762, 1016×1016, and 1270×1270, in which 1016×1016 context mask leads to the best result. In Fig. 8, we demonstrate that our model is able to decrease the boundary artifacts and refine the contour of the semantic mask. Our network can collaborate with prior models. E.g., along with GLNet [4], it can also boost its mIOU from 71.6 to 72.6 on DeepGlobe.

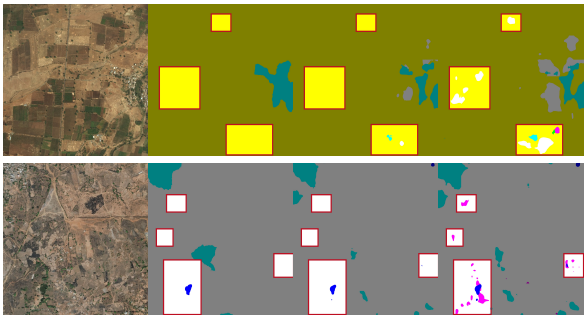
Memory cost and timing performance. During the inference stage, our local segmentation model costs around 3167MB memory for each patch of the images in DeepGlobe and Inria Aerial, which do not increase much computation overhead over FCN-8s (2477MB). As an independent model from the segmentation model, our refinement network costs 1165MB memory. Thus, the memory usage

Fusion method	DeepGlobe	Inria Aerial
Averaging Fusion	72.67	73.29
Weighted Fusion	72.99	73.43
Adaptive Fusion	**73.22**	**73.53**

Table 5. Analysis of feature fusion scheme.

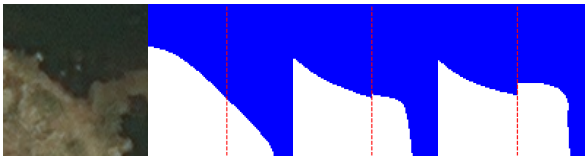
Refinement	Context Size	DeepGlobe	Inria Aerial
Montaging	-	73.22	73.45
Averaging	-	73.22	73.53
Ours	762	73.24	73.63
	1016	**73.45**	**73.66**
	1270	73.36	73.63

Table 6. Context sizes for mask refinement on public datasets.



Input Image Ground-truth w/ Fusion w/o Fusion

Figure 7. Examples show the efficacy of our adaptive fusion.



Input Image Ground-truth w/ Refinement w/o Refinement

Figure 8. The example shows boundary artifacts on adjacent patches, where boundary is represented in dash line.

of our model is comparable to existing semantic segmentation models. For the timing performance, our segmentation model processes each patch in 0.15s and our refinement model requires 0.06s per patch during inference. Overall, for each instance from DeepGlobe and Inria Aerial, our model needs to cost 8s and 26s, compared with FCN-8s (3s and 9s), GLNet (6s and 19s), and CascadePSP (9s and 37s).

5. Conclusion

We introduce a locality-aware contextual correlation based segmentation model to process local image patches. In addition, we present a contextual semantics refinement network that is enabling to reduce the boundary artifacts and refine mask contours during the process of creating the final high-resolution mask.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No. 62072110, 61873067, 61972162); Guangdong International Science and Technology Cooperation Project (No. 2021A0505030009); Guangdong Natural Science Foundation (No. 2021A1515012625); Guangzhou Basic and Applied Research Project (No. 202102021074).

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 39(12):2481–2495, 2017. 2, 6, 7

[2] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2018. 2, 3, 6, 7

[3] L. C. Chen, Y. Yi, W. Jiang, X. Wei, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 3

[4] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *CVPR*, pages 8924–8933, 2019. 2, 3, 5, 6, 7, 8

[5] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *CVPR*, pages 8890–8899, 2020. 1, 2, 6, 7

[6] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *CVPRW*, pages 172–181, 2018. 5

[7] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019. 2

[8] Golnaz Ghiasi and Charless C Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *ECCV*, pages 519–534. Springer, 2016. 3

[9] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015. 3

[10] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *CVPR*, pages 7519–7528, 2019. 2

[11] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, pages 603–612, 2019. 2

[12] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, pages 9799–9808, 2020. 2

[13] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, pages 1925–1934, 2017. 3

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, Oct 2017. 5

[16] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *CVPR*, pages 82–92, 2019. 3

[17] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. In *ICLR*, 2016. 3

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2, 6, 7

[19] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, pages 10034–10043, 2020. 2

[20] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IGARSS*, pages 3226–3229. IEEE, 2017. 5

[21] Davide Mazzini. Guided upsampling network for real-time semantic segmentation. In *BMVC*, 2018. 3

[22] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, December 2015. 2

[23] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. In *CVPR*, 2016. 2

[24] Rudra PK Poudel, Ujwal Bonde, Stephan Liwicki, and Christopher Zach. Contextnet: Exploring context and detail for semantic segmentation in real-time. In *BMVC*, 2018. 3

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 3, 6, 7

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4, 6

[27] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *CVPR*, pages 5229–5238, 2019. 2

[28] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville. Reseg: A recurrent neural network-based model for semantic segmentation. In *CVPRW*, 2016. 3

[29] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 3

[30] Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Yongdong Zhang. Cgnet: A light-weight context guided network for semantic segmentation. *TIP*, 30:1169–1179, 2020. 2

[31] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, pages 648–663. Springer, 2016. 3

[32] Zongxin Yang, Linchao Zhu, Yu Wu, and Yi Yang. Gated channel transformation for visual recognition. In *CVPR*, pages 11794–11803, 2020. 3

[33] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018. 3

[34] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *ECCV*, pages 405–420, 2018. 2, 6, 7

[35] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 3, 6, 7

[36] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *CVPR*, pages 13065–13074, 2020. 3

[37] Peng Zhou, Brian Price, Scott Cohen, Gregg Wilensky, and Larry S Davis. Deepstrip: High-resolution boundary refinement. In *CVPR*, pages 10558–10567, 2020. 1

[38] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019. 3