Article

Affiliated Fusion Conditional Random Field for Urban UAV Image Semantic Segmentation

Yingying Kong ^{1,*} , Bowen Zhang ¹ , Biyuan Yan ^{1,2}, Yanjuan Liu ¹, Henry Leung ³ and Xiangyang Peng ²

¹ Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China; vinzhang@nuaa.edu.cn (B.Z.); yanbiyuan@nuaa.edu.cn (B.Y.); liuyanjuan@nuaa.edu.cn (Y.L.)

² Nanjing Research Institute of Electronics Engineering, Nanjing 210007, China; wwukt@163.com

³ University of Calgary, Calgary, AB T2P 2M5, Canada; Leungh@ucalgary.ca

* Correspondence: yayako_zy@nuaa.edu.cn; Tel.: +86-185-5140-8717

Received: 11 January 2020; Accepted: 10 February 2020; Published: 12 February 2020

Abstract: Unmanned aerial vehicles (UAV) have had significant progress in the last decade, which is applied to many relevant fields because of the progress of aerial image processing and the convenience to explore areas that men cannot reach. Still, as the basis of further applications such as object tracking and terrain classification, semantic image segmentation is one of the most difficult challenges in the field of computer vision. In this paper, we propose a method for urban UAV images semantic segmentation, which utilizes the geographical information of the region of interest in the form of a digital surface model (DSM). We introduce an Affiliated Fusion Conditional Random Field (AF-CRF), which combines the information of visual pictures and DSM, and a multi-scale strategy with attention to improve the segmenting results. The experiments show that the proposed structure performs better than state-of-the-art networks in multiple metrics.

Keywords: semantic image segmentation; CRF; DSM; UAV; remote sensing

1. Introduction

Conditional Random Field (CRF) is a structured prediction model based on graph structures, which is a Markov Random Field (MRF) of random variables Y for given random variables X [1]. CRF has been widely used in various structured prediction tasks, such as estimating scores in Go games [2], separating specific genes from DNA [3], word segmentation in the field of natural language processing [4,5], and image segmentation for computer vision [6,7], etc. Since CRF can utilize lower-level contextual information at the structural level, it is well suited for these structured prediction tasks.

Semantic image segmentation, a pixel-level recognition task labeling the object category to each pixel in the image, is a typical structured prediction task. Semantic image segmentation is the basis of scene-parsing tasks, which have great potential in further applications such as automatic cruise, the landing of drones, and autonomous vehicles. The early application of CRF in the field of semantic image segmentation mainly utilizes a second-order CRF to model an image [8], which reads the associated information of four neighborhoods for each pixel, and uses two types of potential functions to model the conditional probability. The used potential functions are a unary potential function, which is only related to the current pixel feature, and a pairwise potential function, which is associated with pixels within the neighborhood. The precise inference of such methods often requires a large computational cost. Therefore, the maximum posterior estimation of the approximated conditional probability is usually obtained by sampling methods or variational methods.

The four-neighbor connectivity CRF has achieved a degree of success in certain fields, yet it also has inherent drawbacks in the model structure: four-neighbor communications can only consider

the neighboring dependence of adjacent pixels, but not the spatial position. It cannot establish a long-range dependence on two pixels that are far away from each other in spatial position so that satisfactory segmentation results cannot be obtained when there is a large cross area, occlusion, repetition, or deformation in the category of the target. In order to solve this problem, a series of CRFs using high-order potential functions are proposed to improve their performance [9,10], which solve the above problems in certain aspects, but the computational complexity of the model is also significantly increased simultaneously.

To solve the above problems, a fully connected CRF with Gaussian kernels in the feature space is proposed [11], which has achieved a good compromise between speed and precision. However, CRFs that use specific functions are still not robust enough, resulting in complex textures and reduced performance, especially with changes in the proportion of long-range images such as those taken by drones or satellites. In order to better consider the spatial information of long-range pixels in order to reduce error classifications in the image parts with complex texture, we introduce geographical information through a digital surface model (DSM). To get a further improvement of segmentation accuracy, this paper proposes a fully connected Affiliated Fusion CRF (AF-CRF) with DSM, which (1) adopts a fully connected CRF structure to take long-range dependence into account; (2) utilizes the corresponding DSM to comprehend the height information of the region of interest (ROI); and (3) takes multi-scale pyramids for both images and DSM for global feature understanding. Compared to a large amount of manually labeled pictures, DSM can be generated automatically by the pictures taken above the ROI. The extra information of elevation of the interested region is believed to improve the performance of the classifier.

2. Preliminary

CRF is a discriminant undirected graph model. Compared with the corresponding generative model, it directly models the conditional probability. In this section, we simply introduce the theory of CRF and its general models.

2.1. Conditional Random Field

The idea of CRF is to model the conditional probabilities of multiple variables with given observations. Let $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ be the observation sequence and $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ be the corresponding label sequence. The purpose of CRF is to model the conditional probability $P(\mathbf{y}|\mathbf{x})$. $\mathbf{y}$ is a structural variable so that some relationship exists between its components.

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ denote undirected graphs with one-to-one correspondence between the observation sequence $\mathbf{x}$ and its label $\mathbf{y}$. y_v denotes an annotation variable corresponding to node v , and $n(v)$ denotes adjacent nodes of v . If each variable y_v of the graph satisfies the Markov property, then there is

$$P(y_v | \mathbf{x}, y_{V \setminus \{v\}}) = P(y_v | \mathbf{x}, y_{n(v)}), \quad (1)$$

where $V \setminus \{v\}$ denotes all nodes except v .

2.2. General CRF Model

From the view of Equation (1), a CRF adds an observation sequence to a MRF, i.e., a CRF is a MRF given the observations. Similar to MRF, CRF defines the conditional probability distribution with potential functions and the cliques on the graph. According to Hammersley–Clifford theory, the probability distribution between multiple variables can be decomposed into a product of multiple factors based on the cliques in the graph, in which each factor is related only to a single maximum clique of the graph [12]. Specifically, for an observation sequence $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ and its corresponding label sequence $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$, let C denote the set of maximum cliques and let $(\mathbf{x}_C, \mathbf{y}_C)$ denote the variables corresponding to a single maximum clique $C \in C$; then, the conditional distribution $P(\mathbf{y}|\mathbf{x})$ can be written as

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \Psi_C(\mathbf{y}_C, \mathbf{x}_C), \quad (2)$$

where $Z = \sum \prod_{C \in \mathcal{C}} \Psi_C(\mathbf{y}_C, \mathbf{x}_C)$ is the partition function that normalizes $\prod_{C \in \mathcal{C}} \Psi_C(\mathbf{y}_C, \mathbf{x}_C)$.

In particular, a general CRF model is applied to the problem of semantic image segmentation. Consider a random field $\mathbf{X}$ defined on a set of variables $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ with label $\mathbf{y} \in L = \{l_1, l_2, \dots, l_k\}$ and another random field $\mathbf{I}$ defined on another set of variables $\{I_1, I_2, \dots, I_k\}$, in which $\mathbf{I}$ denotes an input image of size N and $\mathbf{X}$ denotes the pixel-level annotation of $\mathbf{I}$. Then, the conditional probability distribution $P(\mathbf{X}|\mathbf{I})$ can be expressed as

$$P(\mathbf{X}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp \left(- \sum_{c \in \mathcal{C}_G} \phi_c(\mathbf{X}_c|\mathbf{I}) \right), \quad (3)$$

where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the graph defined on $\mathbf{X}$. $\mathcal{C}_G$ is the set of all the cliques in $\mathcal{G}$ and every clique c in $\mathcal{C}_G$ introduces a potential function ϕ_c [13]. The Gibbs energy of the annotation $\mathbf{x} \in \mathcal{L}^N$ is

$$E(\mathbf{x}) = \sum_{c \in \mathcal{C}_G} \phi_c(\mathbf{x}_c|\mathbf{I}). \quad (4)$$

Then, the MAP estimation of the CRF is

$$\mathbf{x}^* = \operatorname{argmax}_{\mathbf{x} \in \mathcal{L}^N} P(\mathbf{x}|\mathbf{I}). \quad (5)$$

In a fully connected CRF, let $\mathcal{G}$ be the complete graph of $\mathbf{X}$, in which $\mathcal{C}_G$ is the set of both unary cliques and pairwise cliques; then, the corresponding Gibbs energy is

$$E(\mathbf{x}) = \sum_i \phi_u(x_i|\mathbf{I}) + \sum_{i < j} \phi_p(x_i, x_j|\mathbf{I}). \quad (6)$$

The unary potential $\phi_u(x_i|\mathbf{I})$ is only related to the feature of a single pixel itself, while the pairwise potential $\phi_p(x_i, x_j|\mathbf{I})$ is related to the similarities and differences between every two pixels, which can be expressed as the linear combination of the Gaussian kernels defined in the feature space

$$\phi_p(x_i, x_j|\mathbf{I}) = \mu(x_i, x_j) \sum_{m=1}^M w^{(m)}(x_i, x_j) \mathbf{k}^{(m)}(\mathbf{f}_i - \mathbf{f}_j), \quad (7)$$

where $\mu(x_i, x_j)$ is the label compatibility function to measure the probability that two labels appear at the same time. For instance, the probability of annotation {ship} and {waters} should be larger than one of {plane} and {waters}, intuitively.

3. Affiliated Fusion Conditional Random Fields

In this section, we describe the proposed model applying to semantically segment urban unmanned aerial vehicles (UAV) images. The first is a brief introduction to the theory of multi-scale and attention analysis. Afterwards, the Affiliated Fusion CRF is proposed.

3.1. Multi-Scale Analysis

Multi-scale analysis is a common method of digital image processing, which involves the representation and analysis of images at multiple resolutions. The advantages of this approach are obvious, in which features that cannot be detected at one resolution are often easier to be detected at another resolution. This section uses the classic image Gaussian pyramid as a multi-scale metric.

The image pyramid is a series of images of the pyramid structure obtained by the original image after multiple times of downsampling operation with the same ratio. The original image size and resolution of the bottom layer are the highest, and the resolution of the upper layer is reduced. The images in each layer have different sizes and resolutions. A complete image pyramid has $n + 1$ layers of image. The use of different scale representations of an image can be thought of as adding another dimension to the image. In addition to the conventional positional dimension (x, y) , a dimension for depicting the current number of pyramid layers is added. This structure is also called the scale space.

According to the sampling theorem, only a fine structure that is sampled with less than $1/4$ wavelength can be eliminated by a smoothing filter so that a correct downsampled image can be obtained. From a scale space perspective, this means that reducing the size of the image needs to be done in synchrony with the proper smoothing operation. The smoothing operation of the image can be performed by various low-pass filters, and the image pyramid obtained by Gaussian smoothing filter can be expressed as

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x-x_0)^2 + (y-y_0)^2}{2\sigma^2}\right). \quad (8)$$

3.2. AF-CRF

The unary potential $\phi_u(x_i|\mathbf{I})$ is only related to the feature of a single pixel, while the pairwise potential $\phi_p(x_i, x_j|\mathbf{I})$ is related to the similarities and differences between every two pixels. Therefore, to take advantage of the contextual information in the scale space, we optimize $\phi_p(x_i, x_j|\mathbf{I})$ in the proposed AF-CRF, which can be expressed as below with the parameters of the model, θ , and we hide the image $\mathbf{I}$ from Equation (7) for convenience:

$$\phi_p(x_i, x_j|\theta) = \sum_{m=1}^M \mu^{(m)}(x_i, x_j|\theta) w^{(m)}(x_i, x_j) k^{(m)}(f_i - f_j), \quad (9)$$

where $k^{(m)}(\cdot)$ refers to Gaussian kernel functions and m is the number of them. $k^{(m)}(\cdot)$ can be written as

$$k^{(m)}(f_i - f_j) = \exp\left(-\frac{1}{2}(f_i - f_j)^T (\Sigma^{-1})^{(m)} (f_i - f_j)\right), \quad (10)$$

where the vectors f_i and f_j are the feature vectors of pixel i and pixel j in the feature space, respectively. $w^{(m)}$ is the weight of each kernel, and $\mu^{(m)}(x_i, x_j|\theta)$ is the label compatibility function. Each Gaussian kernel can be defined by the inverse of a covariance matrix $(\Sigma^{-1})^{(m)}$.

The Gaussian kernels of the pairwise potential of the CRF with the DSM mentioned above are constructed with color vector $\mathbf{I}$, position vector $\mathbf{p}$, and height h . The addition of h brings the spatial sensitivity in the pairwise potential of the model, which is defined in the Gaussian kernel as

$$k^{(m)}(f_i - f_j) = w^{(1)} \exp\left(-\frac{|\mathbf{p}_i - \mathbf{p}_j|^2}{2\theta_\alpha^2} - \frac{|\mathbf{I}_i - \mathbf{I}_j|^2}{2\theta_\beta^2} - \frac{|h_i - h_j|^2}{2\theta_\gamma^2}\right) + w^{(2)} \exp\left(-\frac{|\mathbf{p}_i - \mathbf{p}_j|^2}{2\theta_\delta^2}\right), \quad (11)$$

where the first term considers that pixels with similar positions, similar colors and small height difference have higher probability of the same label categories. The second one considers the smoothing terms, which only points out that pixels with similar positions have a higher probability of the same label categories.

As for the label compatibility function $\mu^{(m)}(x_i, x_j | \theta)$ that appeared in Equation (9), a simple and practical one is the Potts model, which can be written as

$$\mu^{(m)}(x_i, x_j | \theta) = I(l_i, l_j), \quad (12)$$

where $I(l_i, l_j)$ is a 0–1 indicator, which is

$$I(l_i, l_j) = \begin{cases} 1, & l_i \neq l_j \\ 0, & l_i = l_j \end{cases}. \quad (13)$$

Although the segmentation result of utilizing the Potts model in CRF is generally acceptable, its disadvantage can be seen from its definition: the Potts model penalizes all inconsistent labels equally. For instance, the penalty of label {ship} and {waters} is equal to that of {plane} and {waters}, which is obviously not intuitive. In order to improve this simple Potts model, we can learn a symmetric adaptive category associated label compatibility function, which is

$$\mu(l_i, l_j | \theta) = \mu(l_j, l_i | \theta). \quad (14)$$

Therefore, a more reasonable penalty is imposed on the inconsistency of the annotation in the semantic segmentation result of the UAV images.

Taking attention into account with inputting the images in S scales to the model and upsampling the output in all scales to the original size, the attention feature $f_{i,c}^s$ is the interest value of the i th pixel belongs to class c on the scale s with the pairwise potential of the CRF, where $s \in \{1, 2, \dots, S\}$. Let $g_{i,c}$ denote the weighted sum of position i and class c of all the attention features of the representation, i.e., $g_{i,c} = \sum_s \delta_i^s f_{i,c}^s$, where the weight δ_i^s for each scale is

$$\delta_i^s = e^{h_i^s} / \sum_{t=1}^S e^{h_i^t}. \quad (15)$$

h_i^t is the output scores of each class (ranges from 1 to C) in each position (ranges from 1 to N). δ_i^s reflects the importance of position i in scale s , which is shared in all channels.

The attention term describes the relationship among the different scales of the CRF's input and combines the differences of focused performance in the different scales with an adaptive weight. With the attention term, the model considers multi-scale information in the inference stage to get a more reasonable prediction.

Assume that the input scales are $s \in \{1, 2, \dots, S\}$ and combine the attention term proposed in the previous section. Considering the pairwise potential as an auxiliary decision condition in the inference stage, the Gibbs energy of the proposed model can be rewritten as

$$E^*(x) = \sum_i \phi_u(x_i | \mathbf{I}) + \sum_s \delta_i^s \sum_{i < j} \phi_{p,s}(x_i, x_j | \mathbf{I}). \quad (16)$$

where $\delta_i^s = e^{h_i^s} / \sum_{t=1}^S e^{h_i^t}$ as shown in Equation (15), and $\phi_{p,s}(\cdot)$ denotes the pairwise potential of the scale s . The new $E^*(x)$ considers the features of color, position, and height, as well as scale, which theoretically helps improve the robustness of the overall model. The workflow is illustrated in Figure 1.

Figure 1. The workflow of Affiliated Fusion Conditional Random Field (AF-CRF).

4. Inference and Learning of the Model

In a fully connected CRF, the bottleneck of computing speed lies in the message-passing step. In order to achieve the fast operation of the proposed AF-CRF, this paper adopt the fast-solving algorithm with the mean field approximation proposed in [11] instead of the precise inference. The model inference is transformed into a Gaussian filtering in the feature space to improve the operation speed and achieve the acceptable accuracy. This section introduces the inference of conditional random fields and the learning methods of their parameters.

4.1. Inference

According to Equation (16), the exact probability distribution can be written as

$$P(\mathbf{X}) = \frac{1}{Z} \tilde{P}(\mathbf{X}) = \frac{1}{Z} \exp \left(\sum_i \phi_u(x_i | \mathbf{I}) + \sum_s \delta_i^s \sum_{i < j} \phi_{p,s}(x_i, x_j | \mathbf{I}) \right). \quad (17)$$

In order to achieve the fast operation of the proposed AF-CRF, the mean field approximation method is used in this section; that is, the probability distribution of the minimum Kullback-Leibler divergence (KL divergence) between the calculation and the accurate probability distribution is approximated instead of calculating the exact probability distribution directly. The approximate probability distribution can be expressed as the product of a series of marginal probability distributions as

$$Q(\mathbf{X}) = \prod_i Q_i(X_i), \quad (18)$$

where $Q_i(X_i)$ is the marginal distribution of each variable in the model. Then, the KL divergence of $Q(\mathbf{X})$ and $P(\mathbf{X})$ can be expressed as

$$\begin{aligned} D(Q||P) &= \sum_x Q(x) \log \frac{Q(x)}{P(x)} \\ &= -\sum_x Q(x) \log P(x) + \sum_x Q(x) \log Q(x) \\ &= -E_{U \sim Q}[\log P(U)] + E_{U \sim Q}[\log Q(U)] \\ &= -E_{U \sim Q}[\log \tilde{P}(U)] + E_{U \sim Q}[\log Z] + \sum_i E_{U_i \sim Q}[\log Q(U_i)] \\ &= E_{U \sim Q}[E(U)] + \sum_i E_{U_i \sim Q}[\log Q_i(U_i)] + \log Z, \end{aligned} \quad (19)$$

where $E_{U \sim Q}$ denotes the expectation in the distribution Q . $Q(x)$ can be derived as a product of independent marginal distribution, which is $Q(\mathbf{X}) = \prod_i Q_i(X_i)$. In this paper, the $Q(\mathbf{X})$ of each scale is inferred separately to focus on the specific scale, which is denoted as $Q_s(\mathbf{X})$. In order to minimize

the KL divergence while ensuring that $Q_s(X)$ and $Q_{s,i}(X_i)$ consist of the proper probability density (i.e., $\int Q_s(X) = 1$ and $\int Q_{s,i}(X_i) = 1$), the following iterative updating formula is used:

$$Q_{s,i}(x_i = l) = \frac{1}{Z_i} \exp \left(-\phi_u(x_i) - \sum_{l' \in \mathcal{L}} \mu(l, l') \sum_{m=1}^K w^{(m)} \sum_{j \in \mathcal{V}} k^{(m)}(f_i - f_j) Q_{s,j}(l') \right). \quad (20)$$

From the view of signal processing, kernel functions can simplify inner product operation in a mapping space. Equation (20) can be expressed as the convolution of Gaussian kernels in feature space, which is

$$\tilde{Q}_{s,i}^{(m)}(l) = \sum_{j \in \mathcal{V}} k^{(m)}(f_i, f_j) Q_{s,j}(l) - Q_{s,i}(l) = \left(G_{(\Sigma^{-1})^{(m)}} \otimes Q_s(l) \right)(f_i) - Q_{s,i}(l), \quad (21)$$

where $G_{(\Sigma^{-1})^{(m)}}$ refers to the Gaussian kernel defined by $(\Sigma^{-1})^{(m)}$. The convolution is represented as a low-pass filter. According to the sampling theorem, the function can be reconstructed from a set of samples whose spacing is proportional to the standard deviation of the filter [11]. Therefore, we can perform convolution by downsampling $Q_s(l)$, using $G_{(\Sigma^{-1})^{(m)}}$ for the convolution, and upsampling the results at the feature points [14].

Truncated Gaussian approximation is a common approximation for Gaussian kernels, in which all values exceeding two standard deviations are set to zero. Since the spacing of samples is proportional to the standard deviation, the truncated Gaussian kernels only support a constant number of sample points. Therefore, by aggregating values from only a constant number of adjacent samples, the convolution can be approximately calculated at each sample.

4.2. Learning

The data-driven parameters that need to be learned are the symmetric label compatibility functions mentioned in previous sections. The symmetry of μ is beneficial to realize the learning algorithm on the one hand and also to the intuition on the other hand: the punishment of the label {waters} for {ship} should be equal to the penalty of the label {ship} for {waters}. In order to effectively learn this symmetric label compatibility function μ , this paper uses the maximum likelihood estimation (MLE) criterion. The goal of MLE is to find a set of parameters that maximize the log-likelihood of the model given the training image I and its annotation result $T^{(n)}$, which is

$$\begin{aligned} l(\mu : T^{(n)}, I^{(n)}) &= \log P(X = T^{(n)} | I^{(n)}, \mu) \\ &= -E(T^{(n)} | I^{(n)}, \mu) - \log Z(I^{(n)}, \mu), \end{aligned} \quad (22)$$

where $I^{(n)}$ is the training image and $T^{(n)}$ is its annotation.

The Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm is adopted to learn the label compatibility functions to maximize the log-likelihood $l(\mu : T^{(n)}, I^{(n)})$. L-BFGS requires gradients to be calculated, which is difficult to estimate accurately because of the calculation of the gradient of the normalization factor Z . Therefore, the mean field approximation described in the previous section is used to estimate the gradient of Z . That is, a simple approximation of the gradient of each training image, which is

$$\frac{\partial}{\partial \mu(a, b)} l(\mu : T^{(n)}, I^{(n)}) \approx -\sum_i T_i^{(n)}(a) \sum_{j < i} k(f_i - f_j) T_i^{(n)}(b) + \sum_i Q_i(a) \sum_{j < i} k(f_i - f_j) Q_i(b), \quad (23)$$

where $T_i^{(n)}(a)$ is a 0–1 indicator. $T_i^{(n)}(a) = 1$ when the annotation of $I^{(n)}$'s i th pixel is a , i.e., $T_i^{(n)}(a) = a$ and $T_i^{(n)}(a) = 0$ otherwise.

5. Experiments and Analysis

In this section, we first introduce the dataset we use and the metric utilized to test the model. Afterwards, some experiments are conducted to test the proposed AF-CRF. The experiments demonstrate that the proposed method increases both the global accuracy (GA) and the Intersection over Union (IoU) and obtains a state-of-the-art result.

5.1. Dataset

The UAV urban images are taken by DJI Phantom 3 Standard at a height of 40 m, and the DSM is generated by PhotoScan [15] at a resolution of 2560×1536 . The image for experiments is resized to the same resolution so that the DSM is manually adjusted to the image with only translation and rotation. Then, we cut the original image and DSM randomly to 160 pieces of 256×256 slices (60 pieces in order without overlapping and 100 pieces at random with overlapping) to train the networks. The image is labeled into five categories: background (not counted in the result), bridge (cls1), road (cls2), sidewalk (cls3), and vegetation (cls4). We randomly select 140 pairs to train and 20 pairs to validate. The origin image, DSM, and the cut slices are shown in Figures 2 and 3, respectively.

Figure 2. The unmanned aerial vehicles (UAV) image and the corresponding digital surface model (DSM) used in the experiment. (a) The UAV image; (b) the corresponding DSM.

Figure 3. The 256×256 slices. (a) UAV cut images; (b) The corresponding DSM; (c) Ground Truth.

5.2. Implement Details

The model is trained with L-BFGS as described above and the loss is per-pixel softmax, which is a multinomial cross-entropy loss in terms of the predicted label x and the ground truth y , which is

$$\text{loss}(x, y) = -\log \frac{\exp(x_{cls})}{\sum_j \exp(x_j)}. \quad (24)$$

Here, x_{cls} is the predicted score of the ground truth class y .

The unary potential adopts the output of some previous works for semantic image segmentation, such as PSPNet-*ds-ss* [16], and sets it as a priori probability that the output is correct, which in this paper is set to 0.5 with a negative logarithm operation. The output score h_i^t in Equation (15) adopts the output of the last layer of PSPNet-*ds-ss*. The label compatibility function is initialized with the Potts model, which performs an identity matrix in the first training stage.

5.3. Results and Discussion

To get a quantitative evaluation result, we adopt global accuracy (GA) and intersection over union (IoU) as metrics, which are

$$GA = \sum_{i=1}^k n_{ii} / \sum_{i=1}^k t_i, \quad (25)$$

$$IoU_{cls} = \frac{n_{ii}}{t_i - n_{ii} + \sum_{j=1}^k n_{ji}}, \quad (26)$$

where t_i is the total number of pixels of class i and the subscript *cls* means the accuracy within the specific class. k is the number of classes and n_{ij} is the number of pixels that belong to class i and were classified to class j . GA represents the performance of the training, while IoU penalizes the false positive classification to demonstrate the performance in semantic segmentation. To get a general evaluation, mean intersection over union (mIoU) is also adopted, which is

$$mIoU = \frac{1}{k} \sum_{i=1}^k \frac{n_{ii}}{t_i - n_{ii} + \sum_{j=1}^k n_{ji}}. \quad (27)$$

Aiming at further applications of semantic segmentation, we also adopted a metric of confidence for the model's outputs, which indicates the probability of the final output category. The confidence of the outputs matters when decisions are made by multi-sensor fusion.

According to Equations (12) and (16), the main parameters of the model are $\theta = \{\theta_\alpha, \theta_\beta, \theta_\gamma, \theta_\delta\}$, $w = \{w^{(1)}, w^{(2)}\}$, and the label compatibility function $\mu(\cdot)$. Due to the model settings, the value of $w = \{w^{(1)}, w^{(2)}\}$ is always near 1 (generally $1 + 0.001$). The smoothing parameters of the position weight has little effect on the experimental results, so it is set here as a fixed value $\theta_\delta = w^{(1)} = w^{(2)} = 1$. According to [11], this paper also set $\theta_\alpha, \theta_\beta = 13$ here. With regard to the height weight parameter θ_γ , when $\theta_\gamma = \theta_\alpha = \theta_\beta = 13$ by grid searching, the model has the best performance, as shown in Figure 4.

Figure 4. Global accuracy varies with the change of θ_γ .

Figures 5 and 6 are the results of eight iterations of the original CRF model and the convergence of KL divergence of the three conditional random fields, respectively, where D-CRF refers to a Dual-CRF (images and DSM) without the multi-scale strategy. As demonstrated, the KL divergence of the models is generally convergent within five iterations; hence, in subsequent experiments, the number of iterations was set to five. In addition, Figure 6 also reveals that the proposed AF-CRF has the fastest convergence speed among the three kinds of CRF.

Figure 5. Outputs of the origin CRF model in every inference iteration, which is nearly converged within five iterations.

Figure 6. Comparison of KL divergence convergence of three CRF models.

Figure 7 is the pyramid structure of image–DSM–output. The higher the pyramid layers, the simpler the original image and the features of the DSM, and the more concentrated the attention of the output results. The output of the three scales is shown in Figure 8. When the scale factor is $1/8$, the attention range is reduced to the extreme, and the output of the whole image is of the same category. On the one hand, it reflects the characteristics of attention in different scales. On the other hand, it also indicates that in practicality, the pyramid scale should not be too large; otherwise, it will lose the relevant features of the region of interest. In subsequent experiments, the number of pyramid scales was only two, i.e., the scale of $1/2$ and the scale of $1/4$.

Figure 7. Multi-scale pyramid demonstration. (a) Original image pyramid; (b) Corresponding DSM pyramid; (c) Output pyramid.

Figure 8. Multi-scale outputs. (a) Scale 1/2; (b) Scale 1/4; (c) Scale 1/8; (d) Synthesis results based on attention.

The overall performance of the proposed AF-CRF is shown in Figure 9. Compared with the output of other models, the output of AF-CRF is the closest to the ground truth, while carefully retaining the output characteristics of the Fully Convolutional Network (FCN) classifier. A quantitative evaluation is shown in Table 1. Since the evaluation of PSPNet-*ds-ss* on several quantitative evaluations has reached a fairly high level, the performance of AF-CRF has only slightly improved. In order to evaluate the performance of AF-CRF better and fairly, a W-5 (Worst-5) index is proposed, which is the improvement of the five worst results of data output after the model, as shown in Table 2. The proposed AF-CRF has significantly improved the disadvantage of the former classifier.

Figure 9. Comparison of the input with output. (a) Original image; (b) Corresponding DSM; (c) Ground truth; (d) Fully Convolutional Network (FCN) output; (e) Original CRF output (with picture input); (f) Original CRF output (with DSM input); (g) Dual-CRF (D-CRF) output; (h) AF-CRF output.

Table 1. The quantitative evaluation of AF-CRF and some other state-of-the-art models. GA: global accuracy, IoU: Intersection over Union.

Model	GA _T (%)	GA _V (%)	IoU _{cls1} (%)	IoU _{cls2} (%)	IoU _{cls3} (%)	IoU _{cls4} (%)	mIoU (%)
DeepLabv3 + -ds-ss [17]	94.62	91.55	79.36	57.36	65.16	59.76	63.16
PSPNet-ds-ss	97.79	96.83	92.67	74.90	81.12	72.14	80.27
DeepLabv3 + -ds-ss + CRF	95.57	93.29	85.45	57.33	69.97	62.31	65.87
PSPNet-ds-ss + CRF	97.75	96.85	93.21	74.71	83.26	72.14	81.02

Table 2. The W-5 index of AF-CRF and some other state-of-the-art models.

Model	GA (%)	IoU _{cls1} (%)	IoU _{cls2} (%)	IoU _{cls3} (%)	IoU _{cls4} (%)	mIoU (%)
DeepLabv3 + -ds-ss	72.37	71.24	43.20	54.81	42.55	53.24
PSPNet-ds-ss	77.28	80.67	61.90	69.14	53.39	67.77
DeepLabv3 + -ds-ss+CRF	85.34	81.47	54.27	60.72	60.55	71.09
PSPNet-ds-ss + CRF	90.97	90.12	70.01	78.21	69.85	78.82

The values of the symmetric label compatibility function of the learned AF-CRF are shown in Figure 10, where the labels are {0: void}, {1: bridge}, {2: road}, {3: sidewalk}, and {4: vegetation}. From Figure 9, the result of learning is that the compatibility between label bridges and roads, as well as roads and sidewalks, is relatively large, while it restrains the label compatibility of roads and pavements and roads and vegetation, which is consistent with the dataset. On the other hand, the imbalance in datasets also affects the learning of label compatibility function, making it more focused on the category of {1: bridge} with the largest number.

Figure 10. The value of the learned label compatibility parameters.

We also test the improvements to some other networks, as shown in Table 3. FCN8s, DeepLab, and U-Net are all classical networks in the field of semantics image segmentation. Their primary output results have been significantly improved after AF-CRF post-processing.

Table 3. The improvements to other networks.

Model	CRF	GA (%)
FCN8s [18]	×	80.3
	√	85.4
DeepLab	×	78.2
	√	82.5
U-Net [19]	×	67.5
	√	79.3

While focusing on the global accuracy and IoU, this paper also considers the confidence of the outputs of the model. As shown in Figure 11, the confidence of the model output is expressed as a thermogram of the probability of the classification results for each pixel in the image. It can be

seen from the figure that the output of AF-CRF not only improves the accuracy of PSPNet-*ds-ss*, but also improves the output confidence of the former FCN classifier, especially in the edge-independent region. Due to the dense label-label connection of the AF-CRF model, the confidence of the outputs in the continuous area of large labeling (as shown in the upper and lower left of Figure 11) has been significantly improved. Confidence has a positive impact on the subsequent application of semantic UAV image segmentation, which to some extent affects the correct decision of subsequent multi-sensor fusion. The quantitative evaluation of the confidence index is shown in Table 4. AF-CRF is better than the PSPNet-*ds-ss* model in both the average confidence and minimum confidence of outputs. More examples of AF-CRF output are shown in Figure 12.

Figure 11. The confidence of model output. (a) PSPNet-*ds-ss*; (b) AF-CRF.

Table 4. The confidence of the model in the outputs.

Metric	PSPNet- <i>ds-ss</i>	AF-CRF
Aver. Conf (%)	97.09	98.65
Min. Conf (%)	35.57	50.04

Figure 12. Examples of AF-CRF output. (a) UAV images; (b) Corresponding DSM; (c) Former classifier outputs; (d) AF-CRF outputs; (e) Ground Truth.

The proposed AF-CRF also has some drawbacks. As shown in the 3rd column of Figure 12, the wrong classification at the top of the image is difficult to correct. Due to the dense connection

between annotations and the influence of the former classifier as a unary potential, AF-CRF inevitably neglects the correct categories that are not appearing in the output of the former one. In addition, as seen in the first column of Figure 12, after AF-CRF, the correct part of the output in the former classifier appears in the wrong classification spots, which is the result of an over-consideration of long-range dependence, and the categories in the location where the errors occur are all so flat that the corresponding DSM fails to provide additional information as well.

6. Conclusions

In this paper, we propose a novel Affiliated Fusion Conditional Random Field (AF-CRF) to semantically segment urban UAV images. The model adopts DSM as an extra geographical information to improve the segmentation result along with a multi-scale analysis with attention. The experiments show that our proposed model improves not only the accuracy but also the confidence of output results, which has considerable benefit to the further applications of urban UAV image semantic segmentation. The limitation of our method is that the DSM of regions of interest must be generated in advance, yet it will be easier to get as the geographic information system develops. The future work will focus on a more specified form of the unary potential.

Author Contributions: Conceptualization, Y.K. and B.Z.; methodology, Y.K. and B.Z.; software, B.Z.; validation, K.Y., B.Z., B.Y., and Y.L.; formal analysis, B.Z.; writing—original draft preparation, B.Z.; writing—review and editing, Y.K., B.Y., and Y.L.; supervision, H.L.; project administration, X.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (No. 61501228); Natural Science Foundation of Jiangsu (No. BK20140825); Aeronautical Science Foundation of China (No.20152052029, No.20182052012); Basic Research (No. NS2015040).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning (ICML), Williamstown, MA, USA, 28 June 2001.
2. Stern, D.H.; Graepel, T.; MacKay, D. Modelling uncertainty in the game of Go. In Proceedings of the 2004 International Conference on Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 13–18 December 2004.
3. Bernal, A.; Crammer, K.; Hatzigeorgiou, A.G.; Pereira, F.C. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Comput. Biol.* **2005**, *3*, 346–354. [[CrossRef](#)] [[PubMed](#)]
4. Taskar, B.; Guestrin, C.; Koller, D. Max-margin Markov networks. In Proceedings of the 2004 International Conference on Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 13–18 December 2004.
5. Sun, X.; Nan, X. Chinese base phrases chunking based on latent semi-CRF model. In Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), Beijing, China, 21–23 August 2010.
6. Martinez, O.; Tsechpenakis, G. Integration of active learning in a collaborative CRF. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 223–228 June 2008.
7. Desmaison, A.; Bunel, R.; Kohli, P.; Torr, P.H.S.; Kumar, M.P. Efficient Continuous Relaxations for Dense CRF. In Proceedings of the 2016 European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 17 September 2016.
8. He, X.; Zemei, R.S.; Carreira-Perpiñán, M.Á. Multiscale conditional random fields for image labeling. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA, 27 June–2 July 2004.
9. Kohli, P.; Kumar, P.K.; Torr, P.H.S. P³ & Beyond: Move Making Algorithms for Solving Higher Order Functions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 1645–1656. [[PubMed](#)]

10. Kohli, P.; Ladický, L.; Torr, P.H.S. Robust Higher Order Potentials for Enforcing Label Consistency. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 24 January 2009.
11. Krähenbühl, P.; Koltun, V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In Proceedings of the 2012 International Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 3–6 December 2012.
12. Li, H. *Statistical Learning Method*; Tsinghua University Press: Beijing, China, 2012.
13. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; MIT Press: Cambridge, MA, USA, 2009.
14. Paris, S.; Durand, F. A Fast Approximation of the Bilateral Filter Using a Signal Processing Approach. *Int. J. Comput. Vis.* **2009**, *81*, 24–52. [[CrossRef](#)]
15. Jaud, M.; Passot, S.; Le Bivic, R.; Delacourt, C.; Grandjean, P.; Le Dantec, N. Assessing the Accuracy of High Resolution Digital Surface Models Computed by PhotoScan and MicMac in Sub-Optimal Survey Conditions. *Remote Sens.* **2017**, *8*, 465. [[CrossRef](#)]
16. Zhang, B.; Kong, Y.; Chen, Y.; Leung, H.; Xing, S. Urban UAV Images Semantic Segmentation Based on Fully Convolutional Networks with Digital Surface Models. In Proceedings of the 10th International Conference on Intelligent Control and Information Processing (ICICIP), Marrakesh, Morocco, 11–16 December 2019.
17. Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
18. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [[CrossRef](#)] [[PubMed](#)]
19. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing & Computer Assisted Intervention (MICCAI), Munich, Germany, 18 November 2015.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).