




## Article

# Automatic Annotation of Airborne Images by Label Propagation Based on a Bayesian-CRF Model

Xiangyu Zhuo <sup>1,2,\*</sup>, Friedrich Fraundorfer <sup>1,3</sup> , Franz Kurz <sup>1</sup>  and Peter Reinartz <sup>1</sup> 

<sup>1</sup> Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany; fraundorfer@icg.tugraz.at (F.F.); franz.kurz@dlr.de (F.K.); Peter.Reinartz@dlr.de (P.R.)

<sup>2</sup> Remote Sensing Technology, Technische Universität München, 80333 Munich, Germany

<sup>3</sup> Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria

\* Correspondence: xiangyu.zhuo@dlr.de; Tel.: +49-8153-28-4235

Received: 4 December 2018; Accepted: 8 January 2019; Published: 13 January 2019



**Abstract:** The tremendous advances in deep neural networks have demonstrated the superiority of deep learning techniques for applications such as object recognition or image classification. Nevertheless, deep learning-based methods usually require a large amount of training data, which mainly comes from manual annotation and is quite labor-intensive. In order to reduce the amount of manual work required for generating enough training data, we hereby propose to leverage existing labeled data to generate image annotations automatically. Specifically, the pixel labels are firstly transferred from one image modality to another image modality via geometric transformation to create initial image annotations, and then additional information (e.g., height measurements) is incorporated for Bayesian inference to update the labeling beliefs. Finally, the updated label assignments are optimized with a fully connected conditional random field (CRF), yielding refined labeling for all pixels in the image. The proposed approach is tested on two different scenarios, i.e., (1) label propagation from annotated aerial imagery to unmanned aerial vehicle (UAV) imagery and (2) label propagation from map database to aerial imagery. In each scenario, the refined image labels are used as pseudo-ground truth data for training a convolutional neural network (CNN). Results demonstrate that our model is able to produce accurate label assignments even around complex object boundaries; besides, the generated image labels can be effectively leveraged for training CNNs and achieve comparable classification accuracy as manual image annotations, more specifically, the per-class classification accuracy of the networks trained by the manual image annotations and the generated image labels have a difference within  $\pm 5\%$ .

**Keywords:** automatic image annotation; label propagation; Conditional Random Field (CRF); Convolutional Neural Network (CNN)

## 1. Introduction

The last decade has witnessed a revolutionary success of deep neural networks. With the support of ever-increasing computing power, various deep neural networks have emerged for a wide range of applications and demonstrated significant improvements compared to traditional machine learning methods. Nevertheless, training the networks requires a large amount of ground truth data. While open image databases like ImageNet [1] and LabelMeFacade [2] are only applicable for specific scenes, manual image annotation is usually inevitable and costs plenty of time and labor. Therefore increasing importance has been attached to the issue of automatic generation of image annotations.

Remote sensing data is often multi-modal, e.g., collected by different sensors (e.g., optical, hyperspectral, radar, LiDAR) and from different platforms (e.g., satellite, airplane, UAV). The ever-growing volumes of remote sensing data convey rich information of the scene from various

scales and resolutions. Meanwhile, the coverage of free geographical information such as open street map (OSM) is expanding rapidly. The way in which the redundant remote sensing data can be fully exploited is raising increasing attentions. In this context, we propose to exploit existing annotated data and incorporate auxiliary information from available remote sensing images to ease the task of image annotation.

As a bridge between aerial and terrestrial photogrammetry, UAV images contribute to comprehensive representation of the scene. Pixelwise classification of UAV imagery is demanded by many applications such as building modeling, however, hand labelling of such a dataset is tedious and time consuming. In view of the fact that aerial images have much larger coverage than UAV images, we seek to propagate the labels from one aerial image to multiple co-registered UAV image. Theoretically, we simply need to annotate one aerial image manually and then transfer the labels to numerous UAV images of the same area, which would dramatically relieve the burden of manual annotation.

In order to tackle the lack of annotated data, a couple of attempts have been made in automatic generation of image annotations. One option is to generate synthetic data. The emergence of synthetic datasets like Virtual KITTI [3] and Synthia [4] reveals its potential in generating image annotations on a large scale, but the quality of synthetic data largely relies on the quality of the generative model. Though such models can find regular patterns of the authentic data, they may not be able to generate distinctive and diverse images like realistic data, which is a severe limitation for most training tasks.

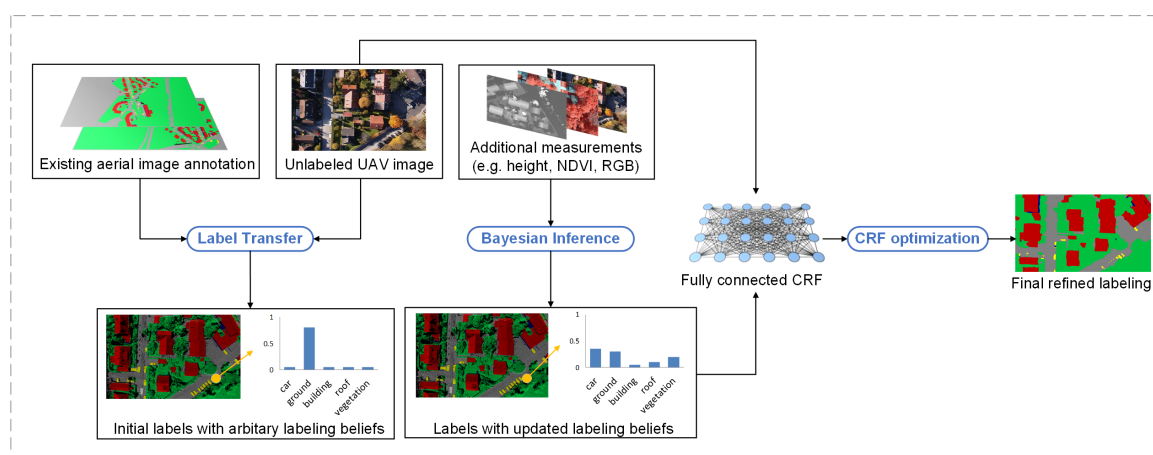
An effective way of automatic image annotation is label propagation, i.e., transfer the pixel labels from source data to target images via certain correspondences. A common case is propagation throughout video frames. Naive methods use optical flow to model the frame to frame propagation [5] but often suffer from occlusion and reappearance of objects. To handle this problem, more sophisticated appearance models have been proposed, such as local shape models in [6], similarity graph in [7] and patch cross-correlation in [8]. Unstructured classifiers, which predict pixel labels independently without neighborhood constraints, mostly use Random Decision Forests [9] and CRFs to get the unaries. While structured classifiers involve neighborhood information to improve classification accuracy. For example, a coupled hidden Markov model (HMM) was proposed in [10] for joint generative modeling of image sequences and their annotation. Joint optimization for temporal motion and semantic labels is used in [11]. A mixture of trees probabilistic graphical model was employed in [8] for temporal association between super-pixels, which can reduce errors caused by short time-window processing in label propagation.

Due to lack of depth information, video-based label propagation approaches are vulnerable to occlusions. In contrast, some methods exploited 3D information as source data for label propagation. For example, the approach presented in [12] exploited 3D information including stereo, noisy point clouds, 3D car models as well as appearance models to generate accurate vehicle classification. Xiao et al. [13] proposed to generate street view image segmentation by projecting 3D annotations onto image superpixels and optimizing the results in a MRF model. However, as the point cloud was generated from Structure from Motion (SFM), this method still had difficulty in dealing with occlusions. In comparison, the hybrid methods presented in [14,15] jointly inferred 2D imagery and 3D point clouds in graphical models and demonstrated improvements in the semantic labeling of both modalities.

Aforementioned methods demonstrate that incorporation of multi-view information can help to improve classification accuracy and temporal consistency of annotations, thus, we are inspired to exploit available radiometric and geometric information for the label propagation task. Nevertheless, existing label propagation methods are generally applied to data of the same source, e.g., across video frames or from terrestrial point cloud to street-view images, where the source data and target imagery have high similarity in view and appearance. However, when it comes to multi-view imagery, the labels propagation suffers from view differences between source data and target images, resulting in sparse and erroneous annotations. For instance, building facades in UAV imagery may be absent or only



partially visible in aerial imagery and thus cannot be annotated, direct label propagation would result in missing labels on the facade area of UAV images. We tackle this problem in a different way: assuming that corresponding nodes in source and target domains have identical labels, we regard the initial propagated labels as prior labeling belief and then update the belief via Bayesian inference when additional measurements (e.g., height, normal vector, NDVI) become available. For instance, observations on NDVI value can be used to distinguish vegetations from the other categories. Towards this goal, we introduce in this paper a Bayesian-CRF graphical model which incorporates available evidence to reason the semantic labels of all image pixels as illustrated in Figure 1. Firstly, we annotate a couple of aerial images and propagate the pixel labels to UAV images by means of geometric transfer, yielding initial labeling for UAV imagery. Subsequently, the pixel labeling probabilities are updated via Bayesian inference given additional measurements such as height information, RGB value or NDVI value. In the end, the pixel labeling is reasoned in a fully connected CRF model defined on the complete set of image pixels, resulting in final refined pixelwise labeling for UAV imagery. The merits of the proposed method lie in the following aspects: (1) the probabilistic essence of our model allows for flexible incorporation of different types of auxiliary information; (2) our method is able to cope with missing labels and the “dragging effect” in optical flow and achieves annotations with high semantic accuracy; (3) our method outperforms manual annotation in the aspect of preserving accurate class boundaries.



**Figure 1.** Workflow of proposed method, including label transfer, Bayesian inference and CRF optimization.

Despite many works on label propagation, few of them have investigated into the possibility of using the propagated labels as pseudo ground truth for training at scale. A notable exception is the work in [16], where a systematic analysis about the quality of pseudo ground truth (PGT) was presented and the impact of PGT on training a CNN was discussed. Similarly, the effect of large scale PGT on deep learning based classification is investigated in [17]. In such cases, the propagated annotations were merely employed as augmented ground truth data and trained together with manually labeled ground truth. By contrast, we demonstrate that the automatic annotations generated by our method can be directly used as ground truth data and achieve comparable accuracy in CNN based classification as manual annotations.

To verify the generalization and robustness of our method, we leverage it for two different applications: (1) UAV image annotation via label propagation from aerial to UAV imagery; (2) aerial image annotation via label propagation from OSM building footprints. In order to explore the effectiveness of deploying the automatically generated image annotations as pseudo ground truth, we train a deep convolutional neural network using these generated annotations for image classification, and compare with classification using manual annotations.

To summarize, the main contributions of this work are:

- We present a novel concept of annotating UAV imagery by transferring the labels from existing annotated aerial imagery.
- We propose a Bayesian-CRF graphical model which can flexibly incorporate 2D and 3D features as auxiliary information, yielding refined image annotations.
- We demonstrate the effectiveness of these annotations as pseudo ground truth data for training deep neural networks for image classification.

The remainder of this paper is organized as follows: Section 2 describes the proposed framework in details. Sections 3 and 4 present the applications of the proposed method in different scenarios and report the experimental results with evaluations. Section 5 interprets the results and describes applicable conditions of the proposed method. Finally, Section 6 concludes the paper.

## 2. Methodology

Label propagation methods transfer labels from annotated source data to a set of target images. In practice, the source annotated data can be aerial imagery, satellite imagery, OSM data, etc. Here we take the case of label propagation from aerial imagery to UAV imagery as an example.

For existing label propagation methods, the source data and target imagery are temporally coherent and similar in view and appearance, e.g., across video frames or from point cloud to images, where the propagation itself can result in accurate image annotations. In our case, the source and target images differ considerably in the aspects of scale, viewing direction, illumination, etc., as UAV imagery is taken at lower altitude with a more oblique view than in aerial imagery. Consequently, the label propagation often results in sparse and noisy annotations. Such missing labels are difficult to identify using only color and texture information, but can be distinguished given appropriate geometric and radiometric information. Intuitively, normal vectors of building facades are generally horizontal while those of the ground are vertical, roofs are expected to be higher than cars, and vegetations appear deeper red in the near-infrared band. The implicit mathematical essence is that different object categories have their distinctive probability distributions over the observed geometric or radiometric data, which can be introduced as evidence in Bayesian theory to update the prior hypothesis and achieve more accurate inference for pixel labeling. To this end, we incorporate these auxiliary evidence from available remote sensing data, and construct a densely connected Bayesian-CRF model to reason about the labels based on evidence.

### 2.1. Model

A CRF can be seen as a Markov Random Field (MRF) globally conditioned on the data. In the context of image classification, a CRF models pixel labels as random variables which have a Markov property and are conditioned upon the image.

More formally, for an input image of size  $N$ , consider a set of random variables  $\mathbf{X} = \{X_1, \dots, X_N\}$  ranging over the image, where  $X_i$  denotes the label assigned to pixel  $i$  and the value is taken from a set of pre-defined semantic labels  $\mathcal{L} = \{l_1, \dots, l_K\}$ . For a global observation (image)  $\mathbf{I}$ , consider a graph  $G = (V, E)$ , where  $V = \{X_1, \dots, X_N\}$ , the pair  $(\mathbf{I}, \mathbf{X})$  can be modeled as a conditional random field which is characterized by a Gibbs distribution  $P(\mathbf{X} = \mathbf{x} | \mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp(-E(\mathbf{x} | \mathbf{I}))$ , where  $\mathbf{x} \in \mathcal{L}$  and  $Z(\mathbf{I})$  is the partition function. Assume the conditional random field  $(\mathbf{I}, \mathbf{X})$  is fully connected, the energy of a label assignment  $\mathbf{x}$  is specified as:

$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i < j} \varphi_p(x_i, x_j) \quad (1)$$

The unary potentials  $\psi_u(x_i)$  encode the probability of a pixel  $i$  taking label  $x_i$ . In our case, pixel-wise label assignment probabilities are initially defined according to the accuracy  $p$  of the transferred weak annotations. For instance, we assume the accuracy of the transferred

annotations to be 80%, i.e., the value of  $p$  is set to 0.8. For an image classification task in five categories  $\{Building, Ground, Vegetation, Car, Clutter\}$ , if a pixel is labeled as *Ground* in the transferred annotation, then the prior label assignment probability for this pixel is represented as a vector of probabilities for each class:  $\{0.05, 0.8, 0.05, 0.05, 0.05\}$ . In case the pixel is not assigned any label, the a priori of label assignment is set to the average value:  $\{0.2, 0.2, 0.2, 0.2, 0.2\}$ . Further, the pixel-wise label assignment probabilities are then updated via Bayesian inference given additional evidence. More specifically, given a sequence of independent and identically distributed evidence features (e.g., height, normal vector, spectral information):  $\mathbf{O} = \{O_1, \dots, O_m\}$ , where  $m$  stands for the number of features. Let  $P(\mathbf{x})$  denote the prior belief of the transferred annotation and  $P(\mathbf{O} | \mathbf{x})$  denote the likelihood of observing  $\mathbf{O}$  given category  $\mathbf{x}$ . In our experiment, the value of  $P(\mathbf{O} | \mathbf{x})$  is empirically defined based on image statistics. The posterior probability of label assignment  $\mathbf{x}$  for given evidence set  $\mathbf{O}$  can be inferred based on Bayes' theorem [18]:

$$P(\mathbf{X} = \mathbf{x} | \mathbf{O}) = \frac{P(\mathbf{O} | \mathbf{x})}{\sum_{\mathbf{x} \in \mathcal{L}} P(\mathbf{O} | \mathbf{x}) P(\mathbf{x})} \cdot P(\mathbf{x}) \quad (2)$$

where,

$$P(\mathbf{O} | \mathbf{x}) = \prod_m P(O_m | \mathbf{x}) \quad (3)$$

Pairwise potentials  $\varphi_p(x_i, x_j)$  encode the cost to assign labels  $x_i, x_j$  to pixels  $i, j$ , respectively, the data-dependent term encourages semantic label coherence of similar pixels. Following the settings in [19], we model the pairwise term as weighted contrast-sensitive Gaussian edge kernels:

$$\varphi_p(x_i, x_j) = \omega_1(x_i, x_j) \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right) + \omega_2(x_i, x_j) \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|c_i - c_j|^2}{2\theta_\beta^2}\right) \quad (4)$$

where  $c_i$  and  $p_i$  respectively represent the color vector and position of pixel  $i$ ,  $\omega_1$  and  $\omega_2$  parametrizes the weights of pairwise features and are both set to 1 in our model. Further,  $\theta_\alpha$ ,  $\theta_\beta$  and  $\theta_\gamma$  control the degree of nearness, similarity and smoothness and are respectively set to 25, 10, 3 in our model.

The pseudo code of the proposed method is listed in Algorithm 1, which explains the Bayesian-based inference in details.

## 2.2. Inference

Basic CRF models only consider neighboring pixels or patches [20,21], which cannot handle long-range connections within the image and often lead to over-smoothing at class boundaries. To solve this problem, higher-order CRFs [22] and hierarchical CRFs [23] have been proposed to integrate region-based hierarchical connections and higher-order potentials, yet the accuracy of such methods largely depend on the accuracy of segmented image regions. By contrast, fully connected CRFs incorporate pairwise potentials for all individual pixels in the image, resulting in significantly higher classification accuracy.

Inference algorithms of CRF models have been widely discussed in many previous research works. The inference problem is basically to find the label assignment  $\mathbf{x}$  with the maximum a posteriori (MAP) of a random field for the given image  $\mathbf{I}$ , i.e.,  $\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{L}} P(\mathbf{x} | \mathbf{I})$ , and can be achieved by minimizing the Gibbs energy function  $E(\mathbf{x})$ . The time for solving the maximization of these marginals is exponential in the size of  $\mathbf{I}$  and thus computationally intractable. To reduce the computational complexity of inference, the mean-field inference algorithm for fully connected CRF models has been proposed in [19], where, the exact CRF distribution  $P(\mathbf{X})$  is approximated with a factorized distribution  $Q(\mathbf{X}) = \prod_i Q_i(X_i)$  that minimizes the KL-divergence  $D(P \parallel Q)$ , the pairwise edge potentials are defined as a weighted sum of Gaussian kernels and the message passing is performed using Gaussian filtering in a Euclidean feature space, which enables highly efficient maximum posterior marginal (MPM) inference. In this paper, we also employ the mean field approximation for the

maximum posterior marginal inference. For implementation, we leverage the framework of Lucas Beyer (<http://graphics.stanford.edu/projects/drfs/>) for construction and inference of the DenseCRF graphical model.

---

**Algorithm 1** Bayesian-CRF Based Image Annotation
 

---

**Input:** Image  $I$ , degree of belief for the transferred annotations  $p$ , number of categories  $K$ , auxiliary evidence set  $O$ , likelihood function  $P(O | x)$

**Output:** Pixel-wise annotation of image  $I$

```

1: procedure UPDATE PIXEL UNARY POTENTIALS BASED ON ADDITIONAL MEASUREMENT
2:   for each pixel  $i$  in  $I$  do
3:      $O \leftarrow$  observation measurement for this pixel.
4:      $P(x_i) = [P_1, \dots, P_K]^T \leftarrow$  prior label assignment probability of this pixel.
5:     if pixel  $i$  is not assigned any label via transferring then
6:        $P(x_i) = [\frac{1}{K}, \dots, \frac{1}{K}]^T$ .
7:     else
8:        $k \leftarrow$  index number of transferred label.
9:       for each  $P_j (j = 1, \dots, K)$  in vector  $P(x_i)$  do
10:        if  $j \neq k$  then  $P_j = \frac{1-p}{n-1}$ 
11:        else  $P_j = p$ 
12:         $P(x_i | O) = \frac{P(x_i) \cdot P(O|x_i)}{P(O)} \leftarrow$  posterior label assignment probability updated by Bayes'
    theorem.
13:
14: procedure SEMANTIC INFERENCE IN CRF MODEL
15:    $E =$  pixel unary potentials + pixel pairwise potentials  $\leftarrow$  Gibbs energy function
16:   Inference by minimizing  $E$ 
  
```

---

### 3. Image Annotation via Label Propagation from Aerial Imagery to UAV Imagery

In this section, we leverage the proposed method to generate pseudo ground-truth data for training. The experiments are comprised two aspects: (1) image annotation via label propagation from aerial imagery to UAV imagery, (2) training a CNN using the generated annotations. We describe data acquisition, introduce experiment settings, present and evaluate the results.

#### 3.1. Data Description

As shown in Figure 2, this dataset consists of two subsets: **Area 1** was acquired over Eichenau, a small village in Germany. It is characterized by dense anthropogenic structures such as traditional-style houses, dense vegetations, roads and moving cars; **Area 2** was acquired over a nearby village Unterrothenstein, including a few detached buildings along the roadside.

Image data were taken on 2nd November 2015. In particular, UAV imagery was captured by a Sony Nex-7 camera at an altitude of 100 m above ground with a slightly oblique view. The average Ground Sampling Distance (GSD) of UAV images is 1.8 cm and the image size is  $6000 \times 4000$  pixels; aerial imagery was acquired by the DLR 4k sensor system [24] at an altitude of 600 m above ground, including two Canon EOS-1DX cameras with  $15^\circ$  sideways looking angle and a FOV of  $75^\circ$  across. The aerial imagery has an average GSD of 20 cm and size of  $5184 \times 3456$  pixel. **Area 1** is covered by both aerial imagery and UAV imagery while **Area 2** is only covered by UAV imagery. Detailed characteristics of the datasets are listed in Table 1.



**Figure 2.** Location of two survey sites **Area 1** (Eichenau) and **Area 2** (Unterrothenstein), Germany.

**Table 1.** Characteristics of the datasets used in the experiment.

Imagery	Date	Size (Pixels)	Height (m)	GSD (cm)	Pitch (°)
Aerial	11/2015	5184 × 3456	600	8.4	15
UAV	11/2015	6000 × 4000	100	1.8	10

We defined six classes for this dataset, i.e., *Building*, *Roof*, *Ground*, *Vegetation*, *Car* and *Clutter*. Where, class *Building* refers to building facades; class *Roof* refers to roofs (including overhangs); class *Ground* refers to bare grounds and roads; class *Vegetation* includes trees, bushes and grassland; class *Car* includes all types of vehicles; the rest categories and indistinguishable objects belong to class *Clutter*. The color coding for labels is illustrated in Figure 3.

From **Area 1**, two aerial images were manually annotated as source-data for label propagation, costing about 60–90 min per frame; in order to compare with the automatically inferred annotations, we manually labeled 28 UAV images as ground-truth data, costing about 20–30 min per frame, and then split them into 23 training samples and 5 testing samples. **Area 2** is only used for testing.



**Figure 3.** Color coding used for label propagation from aerial to UAV imagery.

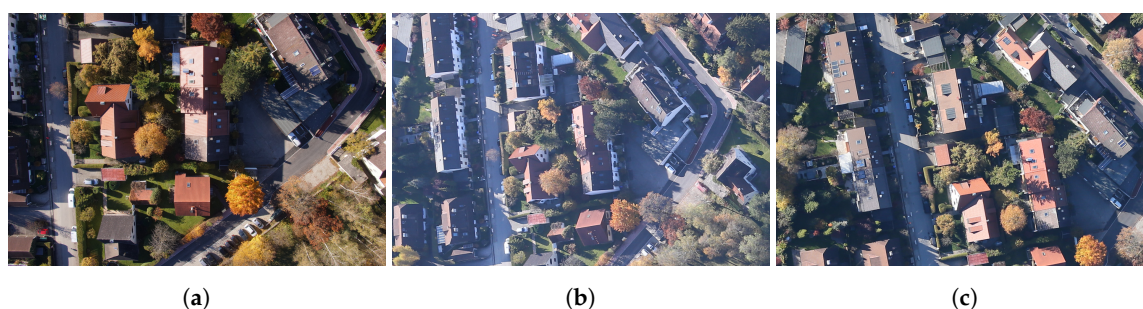
### 3.2. Data Pre-Processing

A high co-registration accuracy is vital to the label propagation between multi-source image data. As the UAV images in our dataset exhibit a much lower geolocalization accuracy than aerial images, we adopted the approach proposed in [25] for co-registration between UAV and aerial images. In short, the method assumes that the aerial images are geo-referenced and have common overlap with UAV images. First, the camera poses of sequential UAV images are solved via Structure From Motion (SFM), and then the nadir UAV images are matched with the aerial images using the proposed matching scheme and generate thousands of reliable image correspondences. Given accurate camera poses of the aerial images, the 3D coordinates of those common image correspondences can be calculated via image-to-ground projection. These 3D points are then adopted to estimate the camera poses of the corresponding nadir-view UAV images. In the end, those UAV images with known camera poses are involved in a global optimization for camera poses of all UAV images. In this way, all UAV images are co-registered to the aerial images with pixel-level registration accuracy. Afterward, we reconstruct



UAV point cloud and Digital Surface Model (DSM) using software *Pix4Dmapper Pro* (version 4.0.25), and then generate a heightmap for each UAV image by deriving heights from the DSM.

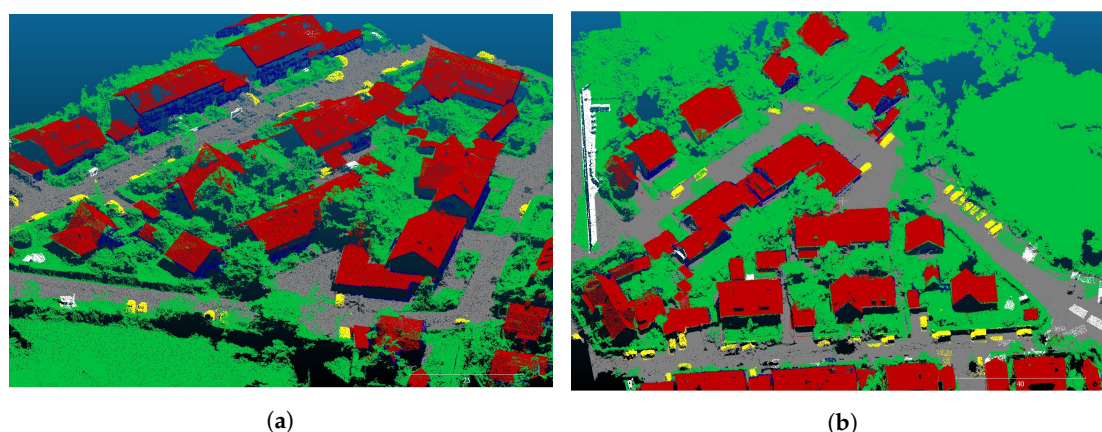
Annotated pixels in aerial imagery can be transferred to UAV imagery based on their orientation parameters. However, individual aerial imagery does not present the same scene as UAV imagery due to their temporal difference and differences in viewing direction, scale, resolution and illumination, etc. For instance, building facades in UAV images maybe not or only partially visible in aerial images. Thus we labeled two aerial images, one left-view and one right-view, to achieve more complete representation of the scene. Figure 4 depicts an oblique UAV imagery and the corresponding region in left-view and right-view aerial images. It can be seen that the combination of the two views can compensate for the view difference to some extent.



**Figure 4.** Comparison of aerial imagery and UAV imagery. (a) UAV image, (b) corresponding region in left-view aerial image, (c) corresponding region in right-view aerial image.

### 3.3. Label Transfer

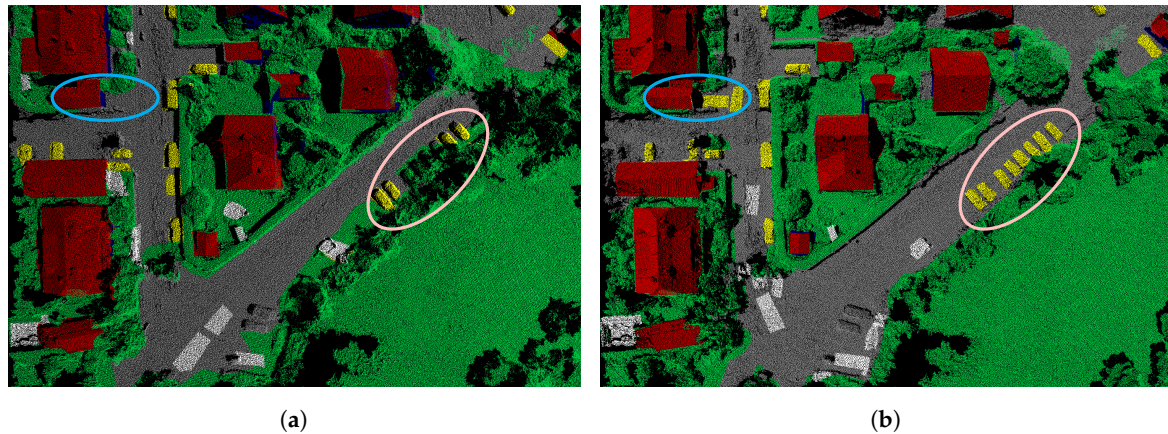
As explained in Section 2, we use the UAV point cloud for geometric label transfer. To be more specific, we project all the 3D points into a labeled aerial image, thus all the non-occluded points get labeled, and then we project these 3D points into UAV images, transferring their labels to corresponding image pixels. Figure 5 illustrates the labeled UAV point cloud, where, Figure 5a shows labels transferred from the left-view aerial image and Figure 5b shows labels transferred from the right-view aerial image. It can be seen that the combination of two aerial images contributes to more enriched and complete representation of the scene.



**Figure 5.** Annotated UAV point cloud with labels transferred from aerial images. (a) labels transferred from the left-view image, (b) labels transferred from the right-view image.

It has to be noted that there are slight temporal differences between the two aerial images, therefore a common 3D point in two labeled point clouds may carry different labels. Besides, occlusions and manual annotation errors also lead to label inconsistencies between two point clouds. Figure 6 illustrates a UAV image with labels transferred from the two labeled point clouds, where examples of label inconsistencies are highlighted.

In view of the fact that one pixel in the initial transferred labeling may carry no label, wrong label, or multiple labels, we refine the pixelwise label assignments in an optimization step using the proposed Bayesian-CRF model. More specifically, the Bayesian inference takes additional measurements (e.g., height) into account to update the pixel labeling beliefs, and then the fully connected CRF model exploits the dense connectivity at the pixel level to reason about the final label assignments, yielding refined labeling for all pixels in the image.



**Figure 6.** A UAV image with labels transferred from two-view aerial images. (a) labels transferred from left-view aerial image, (b) labels transferred from right-view aerial image. Highlighted areas indicate label inconsistency.

### 3.4. Inference

#### 3.4.1. 3D Point Unary Potentials

The 3D point unary potentials can be derived from either a hard manual labeling or a probability distribution computed by a pixel-wise classifier such as MRF or the softmax function of a CNN. In our case, we obtain the unary potential of each 3D point from the labeling of the point cloud.

More formally, let  $\mathbf{P}$  denote the set of non-occluded 3D points in the input UAV point cloud and  $s_i$  denote the label assigned to each point  $i \in \mathbf{P}$ . The domain of each variable  $s_i$  is a set of labels  $\mathcal{L} = \{l_1, \dots, l_K\}$ , where  $K$  denotes the number of classes. In our case,  $K = 6$ ,  $\mathcal{L} = \{Building, Roof, Ground, Vegetation, Car \text{ and } Clutter\}$ .

It needs to be noted that the labeling of UAV point cloud can be transferred from multiple labeled aerial images. Assume we project all non-occluded points of the point cloud into  $n$  annotated aerial images to transfer labels (in our experiment two aerial images are annotated, i.e.,  $n = 2$ ), each 3D point is therefore assigned with  $n$  sets of labels and the corresponding prior label assignment probabilities are denoted by  $P(s_i^1), \dots, P(s_i^n)$ . In order to combine the information from multiple views, we fuse the potentials by taking the average value

$$P(s_i) = \frac{\sum_{j=1}^n P(s_i^j)}{n} \quad (5)$$

where, the initial value of pixelwise label assignment probability  $P(s_i^j)$  is set according to the belief in the manual annotation of aerial images, as described in Section 2.1.

#### 3.4.2. Pixel Unary Potentials

Pixel unary potential encodes the probability of a image pixel  $i$  taking label  $x_i$ . We project all the labeled 3D points of the UAV point cloud into UAV images to transfer the labels. The labels of point cloud are previously transferred from annotated aerial images which are slightly misaligned with UAV

images, besides, the point cloud has higher spatial resolution than UAV images. As a result, multiple 3D points carrying different labels may be projected to a same pixel on the UAV image, i.e., one pixel may be assigned multiple labels. For a pixel  $i$  on the UAV image, let  $\{P(s_1), \dots, P(s_m)\}$  denote the set of the prior probabilities of corresponding 3D points, where  $m$  denotes the number of 3D points which are projected onto this pixel. The prior probability of a image pixel  $i$  taking label  $x_i$  is assigned the average a priori of corresponding 3D points, i.e.,

$$P(x_i) = \frac{\sum_{j=1}^m P(s_j)}{m} \quad (6)$$

where,  $P(x_i)$  is namely the prior belief in Equation (3).

### 3.4.3. Model Parameter Settings

Given additional evidence, the label assignment probabilities are then updated using the Bayesian theory. We employ in this experiment the relative height above the ground,  $H$ , as the evidence, which was obtained by ground filtering using the Top-hat algorithm [26]. As an instance of  $\mathbf{O}$  in Equation (3),  $H$  in this experiment exclusively refers to the observations of height. The likelihood of height measurement for given class  $\mathbf{x}$  is denoted by  $P(H | \mathbf{x})$ . Considering the fact that height values are continuous,  $P(H | \mathbf{x})$  is specified as the probability density function of  $H$  for given class and modeled as normal distribution for the sake of simplification. Since the class *building* (facades) has no height measurements on heightmaps, their prior labeling probabilities were not updated via Bayesian inference. In some cases, a class includes several sub-classes which have considerably different height distributions, e.g., *Vegetation* is comprised of trees and grasslands, we then model the likelihood function as a weighted sum of normal distributions of the sub-classes, i.e.,  $f(H; \mu, \sigma^2) =$

$$\sum \omega_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(H-\mu_i)^2}{2\sigma_i^2}}.$$

Values of these hyperparameters (e.g.,  $\omega, \mu, \sigma$ ) are empirically set based on the manual estimation of image statistics. More specifically, we inspect a DSM of the surveyed area to obtain an overview of the general height distribution of each object category, an example of the likelihood functions is visualized in Figure 7, where the class *roof* has a lower bound of height of 2 meters as roofs are generally higher than that. Then the values of these parameters are estimated to fit such distribution. Furthermore, we can manually fine-tune the hyperparameters during the experiment by inspecting the inference output. An example of hyperparameter settings is listed in Table 2. It needs to be noted that the parameters listed in the table are merely for the sake of understanding and repeating the experiments, but according to our experience, the performance of inference is not sensitive to the parameter setting as long as it reasonably approximates the actual distribution.

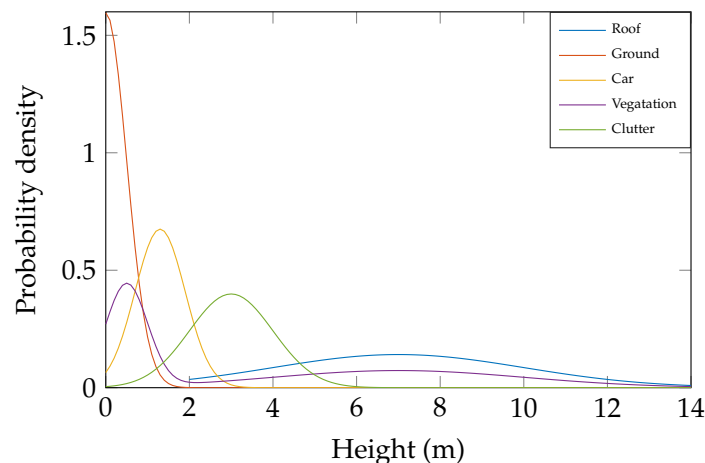


Figure 7. Probability distribution of height for Area 1.

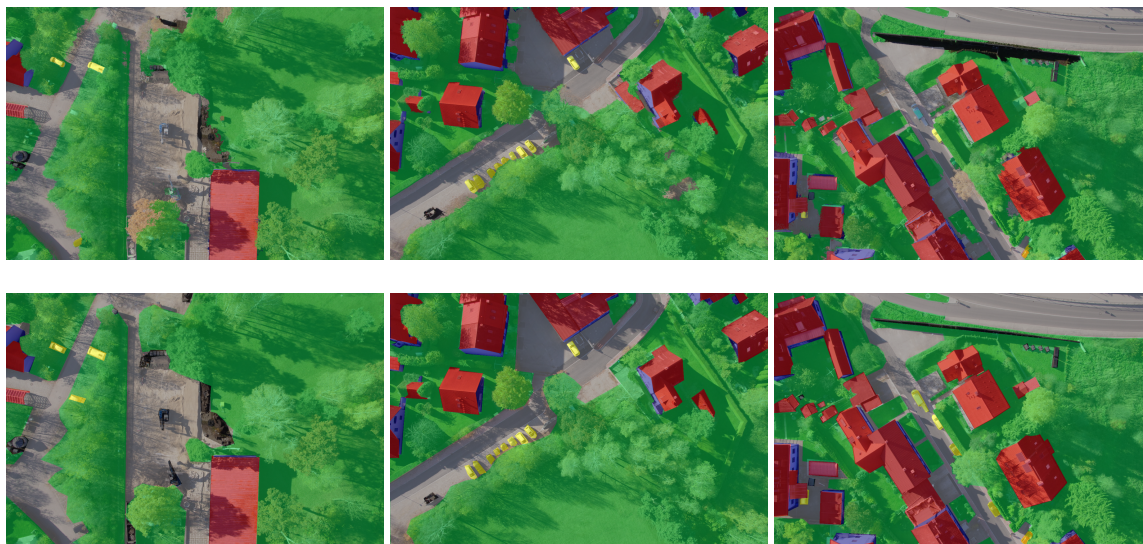


**Table 2.** Parameter settings of probability distribution functions for height, Eichenau Dataset.

Parameter	Ground	Roof	Car	Clutter	Vegetation
$\omega$	2.0	1.0	1.0	0.5	1.0, 0.5
$\mu$	0	7.0	1.3	0.5	3.0, 7.0
$\sigma$	0.5	3.0	0.6	0.5	1.0, 3.0

### 3.4.4. Inference

We performed inference of the CRF model based on the implementation (<https://github.com/lucasb-eyer/pydensecrf>) of [19]. In Figure 8, the top row depicts a few examples of automatically generated annotations while the bottom row shows corresponding manually labeled annotations. It can be seen that the inferred annotations have high semantic accuracy and conform well to the image gradients at class boundaries. On the other hand, there are also a few errors in the inferred annotations, which are caused by three major factors: (1) low contrast in dark or shaded areas; (2) strong gradient at shadow borders; (3) wrong height value (especially for moving cars).



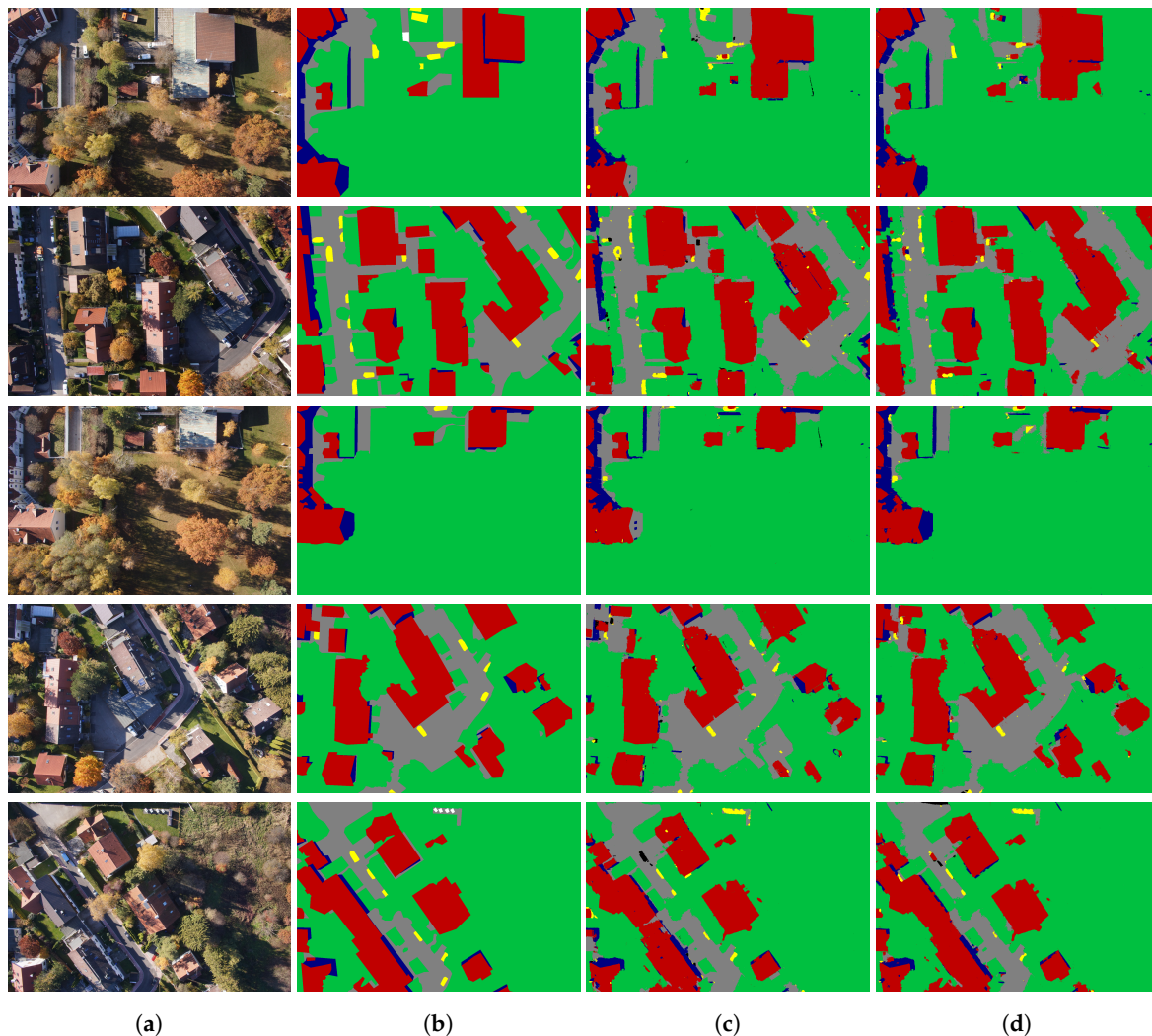
**Figure 8.** Comparison of inferred image annotations (**top row**) and manual annotations (**bottom row**) overlaid on original UAV images.

### 3.5. Training a CNN Using Generated Annotations

In order to validate the utility of the automatically generated annotations, we deployed them as ground-truth data to train a fully convolutional network (FCN) [27] for image classification, and compared the performance with the classification using manual annotations. To be specific, we selected 28 UAV images featuring different regions of the scene and manually labeled them as ground-truth data, which are then split into 23 training samples and 5 testing samples. In parallel, we applied the proposed method to annotate the 23 images as pseudo ground-truth data for training. Both sets of training data are augmented via cropping and rotating, resulting in 8208 images with the size of  $300 \times 300$  pixels.

Figure 9 shows the predictions on test data using manual and automatic ground-truth data. Where, Figure 9a shows the UAV images for testing, Figure 9b lists the corresponding ground truth, Figure 9c illustrates the predictions using manually labeled training data and Figure 9d shows the predictions deploying automatically generated training data. We employ the Intersection over Union (IoU) as the metric for evaluating classification accuracy. Let  $n_{ij}$  be the number of pixels of class  $i$  predicted to belong to class  $j$ , then IoU of class  $i$  is defined as  $\frac{n_{ii}}{\sum_j n_{ij} + \sum_j n_{ji} - n_{ii}}$ , namely the area of overlap divided by the area of union. Table 3 lists the IoU values of each class. It can be seen that

the automatically generated training data achieved comparable classification accuracy as manually labeled training data for static classes such as *Roof*, *Building* and *Ground*, this is due to the fact that the inferred annotations have higher semantic accuracy around class boundaries. However, manual annotations outperformed the inferred annotations for class *Car* due to the existence of moving cars. First, a number of cars in the UAV images are not pictured in source images, which leads to wrong priors after label propagation; second, the heightmap of UAV images has inaccurate height values for a couple of moving cars, which yield wrong evidence in Bayes inference. In summary, erroneous prior and unreliable evidence often lead to wrong inference.



**Figure 9.** Comparison of predictions on Area 1. (a) original UAV images, (b) corresponding ground-truth, (c) predictions using manually annotated training data, (d) predictions using automatically generated training data.

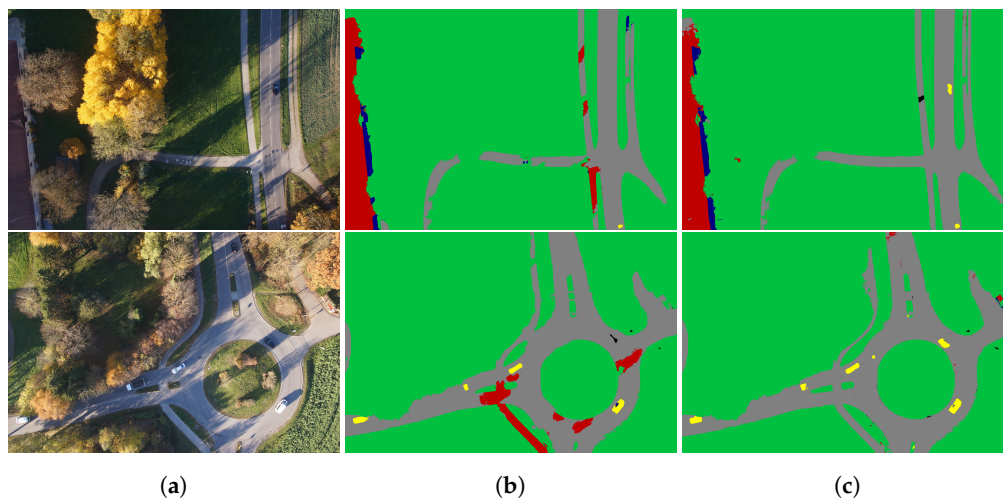
**Table 3.** Comparison of classification accuracy (IoU) using automatic training data and manual training data.

	Roof	Building	Veg	Car	Ground
Manual	84.43	56.47	94.17	39.74	74.47
Inferred	87.48	61.93	93.79	36.21	78.09



### 3.6. Generating Image Annotation on a Scale

The superiority of the proposed method lies also in its ability to generate image annotations on a scale. Due to the high expense of time and labor, the quantity of manually labeled ground-truth data is limited and sometimes not enough to achieve reasonable classification performance. By contrast, our method can generate image annotations on a large scale exempt from manual work, ensuring sufficient amount of training data. To verify the impact of training data quantity, we applied the network trained with the manual annotations (23 image frames) for image classification in **Area 2**. The predictions, as shown in Figure 10b, demonstrate deficient accuracy. In contrast, we generated annotations for 72 image frames of **Area 1** by automatic inference, and then used them to train a CNN. Afterwards, we tested the trained network for image classification in **Area 2**. The prediction results, as depicted in Figure 10c, demonstrate apparently better semantic accuracy.



**Figure 10.** Comparison of predictions on **Area 2**. (a) original UAV images, (b) predictions using 23 manually labeled frames from **Area 1** as training data, (c) predictions using 72 automatically labeled frames from **Area 1** as training data.

## 4. Automatic Image Annotation via Label Propagation from OSM Footprints to Aerial Imagery

In the last section, we have demonstrated the feasibility of automatic image annotation via label propagation, yet manual labeling of source images is still required. In this section, we leverage various remote sensing data for fully automatic image annotation. In particular, we propagate OSM footprints to aerial imagery to generate initial labeling for buildings. Then the initial imprecise labeling is improved by the proposed Bayesian-CRF model, which takes additional measurements (height and NDVI) into account to update multi-class labeling beliefs and exploits the dense connectivity at pixel level for labeling refinement. The experiment is tested on the ISPRS Vaihingen dataset.

### 4.1. Data Description

The Vaihingen dataset was provided by the German Association of Photogrammetry and Remote Sensing (DGPF) [28]. The images were captured in the summer of 2008 over Vaihingen, a medium-size village in Germany. The survey site is characterized by various buildings, including small detached houses, traditional buildings with complex shapes and high-rising buildings surrounded by trees.

Image data used for this experiment includes 33 patches of true orthophoto (TOP) with a GSD of 9 cm, each accompanied by a corresponding DSM with the same spatial resolution. The orthophotos were generated by Trimble INPHO OrthoVista as 8 bit TIFF files with three bands, i.e., near infrared, red and green bands. As the relative height above the ground is more interesting for our experiment rather than the absolute elevation, we generated heightmaps for each TOP by filtering and interpolating the

ALS point cloud using LASTools (<https://github.com/LAStools/LAStools>) The generated heightmaps have a grid size of 9 cm, the same as the TOPs.

There are 16 available ground-truth annotations of the TOPs, which are manually labeled into six categories, i.e., *Impervious surfaces*, *Building*, *Low vegetation*, *Tree*, *Car* and *Clutter*. In our implementation, we kept classes *Building*, *Low vegetation* and *Tree*, and merged the rest categories (*Impervious surfaces*, *Car* and *Clutter*) into a new category *Ground*. Corresponding color coding is illustrated in Figure 11. We selected twelve labeled TOPs (areas: 1, 3, 5, 7, 11, 15, 21, 26, 28, 32, 34 and 37) for training and the rest four labeled TOPs (areas: 13, 17, 23 and 30) for testing.

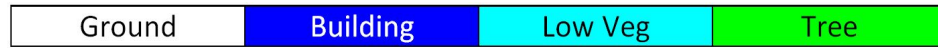


Figure 11. Color coding used for Vaihingen dataset.

The OSM footprint data used in experiments was downloaded on 21 May 2018. Despite the large time offset compared with imagery data, the OSM footprints still have sufficient completeness in most areas. According to manual inspection, the position accuracy of OSM footprints ranges from a few decimeters in inner city to several meters in rural areas.

## 4.2. Automatic Image Annotation

### 4.2.1. Pixel Unary Potentials

The pixel unary potential is derived from the OSM footprints data only. More formally, let  $I$  denote the set of pixels in TOP and  $\mathcal{L}$  denote the set of  $K$  pre-defined labels. Projecting OSM building footprints into a TOP image, the corresponding projection area is denoted by  $I_b$  ( $I_b \subseteq I$ ) and a degree of belief for the OSM footprints is denoted by  $p$ . Let  $l_b$  denote the label index of class *Building*,  $P(x_i)$  encodes the prior belief of a pixel  $i$  ( $i \in I$ ) taking the label  $x_i$  ( $x_i \in \mathcal{L}$ ), which is defined as:

$$\begin{aligned} \forall i \in I_b, P(x_i) &= \begin{cases} p & x_i = l_b \\ \frac{(1-p)}{K-1} & x_i \neq l_b \end{cases} \\ \forall i \notin I_b, P(x_i) &= \begin{cases} 1-p & x_i = l_b \\ \frac{p}{K-1} & x_i \neq l_b \end{cases} \end{aligned} \quad (7)$$

In our case,  $\mathcal{L} = \{\text{Building}, \text{Low vegetation}, \text{Tree}, \text{Ground}\}$ ,  $K = 4$ ,  $p = 0.7$ .

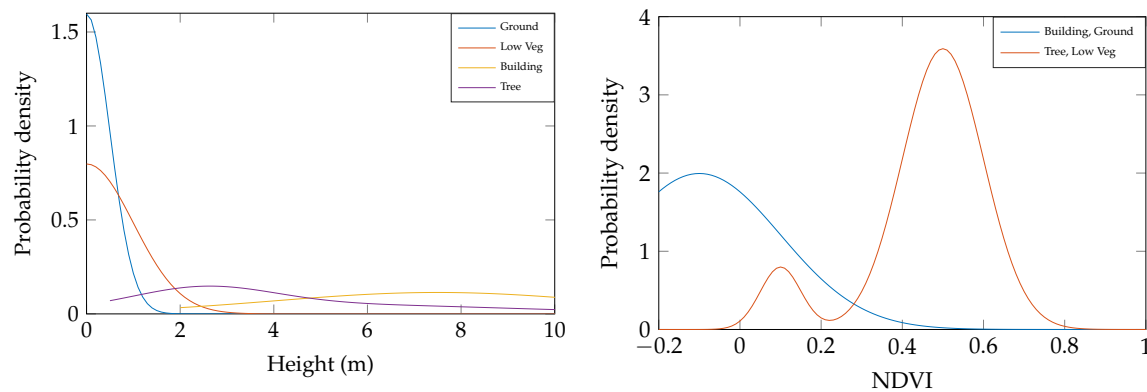
Afterwards, the prior label assignment probability is updated via Bayesian inference given additional evidence. In this experiment, we employ the height and NDVI values as additional evidence, denoted by  $H$  and  $N$ . Namely, the parameter  $\mathbf{O}$  in Equation (3) is the evidence set  $\{H, N\}$  in this experiment. Particularly, the height value is extracted from the heightmap and the NDVI is calculated based on image radiometric information using following formula:

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)} \quad (8)$$

where NIR and Red are the gray values of the TOP tiff files.

The likelihood of height for given class  $\mathbf{x}$ , denoted by  $P(H | \mathbf{x})$ , is modeled as a normal distribution or weighted sum of multiple normal distributions, i.e.,  $f(H; \mu, \sigma^2) = \sum \omega_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(H-\mu_i)^2}{2\sigma_i^2}}$ . As illustrated in Figure 12, the lower bounds for the height of *Tree* and *Building* were set as 0.5 m and 2 m. Values of these function parameters are empirically estimated in the same way as described in Section 3.4.3. Table 4 listed the parameter settings used in our experiment. The multiple parameter settings for class *Tree* is due to the fact that the pre-defined class *Tree* is comprised of bushes and tall trees which have different height distributions, therefore we model its likelihood function as a weighted sum of normal distributions of the sub-classes.

Similarly, the likelihood of NDVI for given class  $\mathbf{x}$ , denoted by  $P(N | \mathbf{x})$ , is illustrated in Figure 12. It needs to be noted that we assume *Tree* and *Low Veg*, *Ground* and *Building* have the same likelihood functions for the sake of simplification. The parameters are listed in Table 5. Considering the fact that the NDVI of vegetations (namely class *Tree* and class *Low Veg*) has relatively lower value in shadows, the likelihood function for NDVI is composed of two normal distributions representing the NDVI in normal cases and in shaded areas.



**Figure 12.** Probability distributions of height and NDVI for Vaihingen dataset.

**Table 4.** Parameter settings of probability distribution functions for height, Vaihingen Dataset.

Parameter	Ground	Building	Low Veg	Tree
$\omega$	2.0	1.0	2.0	0.4, 0.5
$\mu$	0	7.5	0	2.5, 5.0
$\sigma$	0.5	3.5	1.0	1.5, 4.0

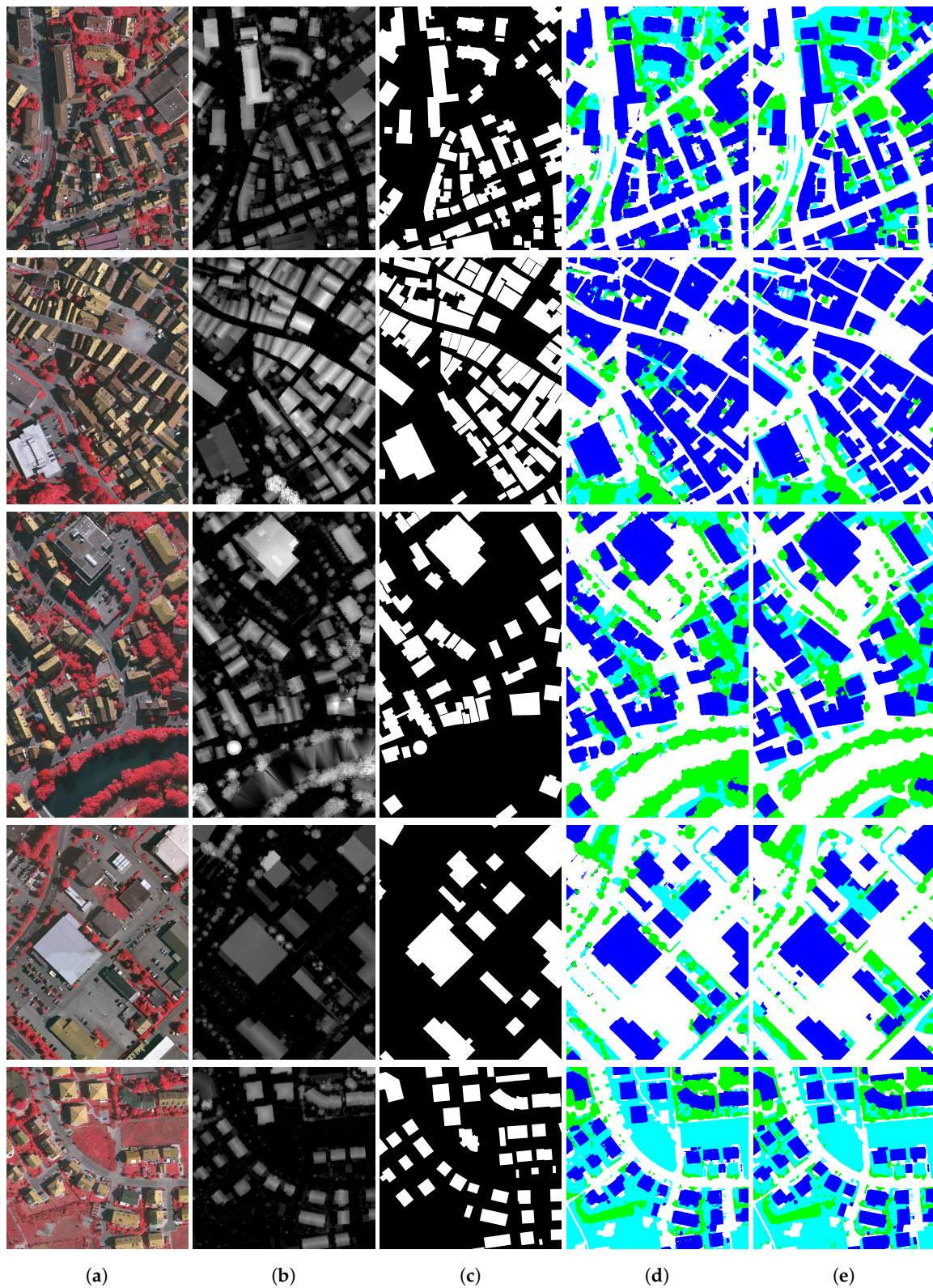
**Table 5.** Parameter settings of probability distribution functions for NDVI, Vaihingen Dataset.

Parameter	Ground & Building	Low Veg & Tree
$\omega$	1.0	0.9, 0.1
$\mu$	−0.1	0.5, 0.1
$\sigma$	0.2	0.1, 0.05

Provided observations on height and NDVI, the posterior distribution  $P(\mathbf{x} | H, N)$ , can be calculated based on Equations (2) and (3). It needs to be pointed out that the parameters listed above are simply an exemplar parameter setting which can achieve effective inference, but in practice the values do not have to be exactly the same. Intuitively, the height evidence helps to distinguish high objects like buildings and trees from low objects like low vegetation and the ground, while the NDVI evidence can effectively differentiate low vegetation and trees from non-vegetation. The combination of height evidence and NDVI evidence yields highly discriminative cues which can tolerate a range of reasonable values.

#### 4.2.2. Inference

A few examples of the inferred annotations are depicted in Figure 13. The source data required for inference is listed in columns Figure 13a–c: (a) true orthophotos with three bands: near infrared, red and green, (b) corresponding heightmaps, (c) building masks projected from OSM footprints. After inference using our Bayesian-CRF graphical model, the generated image annotations are depicted in column (d), and the manually labeled ground-truth data is listed in column (e). It can be seen that the automatically generated annotations have high similarity with manual annotations in general. However, in dark or shaded areas, there are a few errors in categories *Low Vegetation* and *Tree*. This is because the NDVI value cannot well distinguish the vegetation from the ground in shaded areas.



**Figure 13.** Comparison of manual annotations and inferred annotations. (a) true orthophoto, (b) corresponding heightmap, (c) building mask projected from OSM building footprint, (d) automatically inferred image annotations, (e) manually labeled ground-truth.



### 4.3. Analysis of Inferred Annotations

It needs to be noted that manual image annotations are not always correct as they are subjective to human bias [29], especially for objects with complicated shapes. In comparison, our method utilizes additional geometric and radiometric information as discriminative cues to reason about the pixel labeling, contributing to higher semantic accuracy of image annotations. Besides, the inferred annotations inherently conform to image gradients and therefore have higher shape accuracy for irregular objects.

#### 4.3.1. Comparison with Manual Annotations

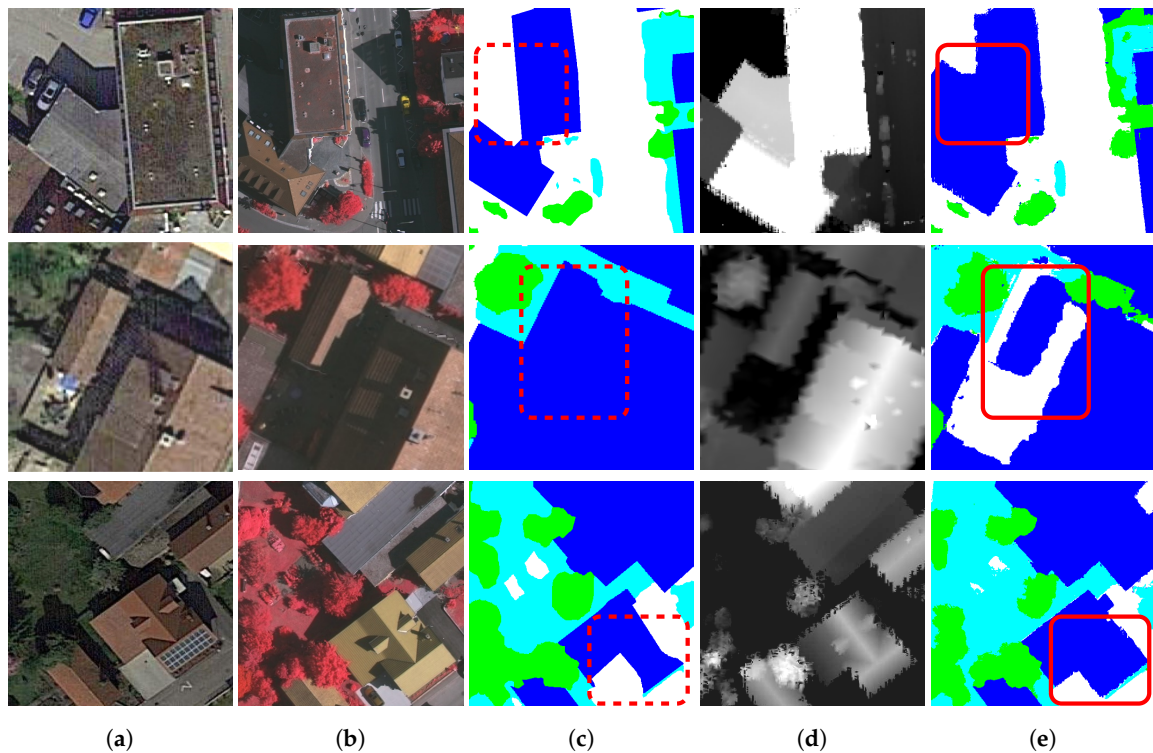
The accuracy of image annotations is mainly affected by objects discrimination and boundaries delimitation [30]. For the former, human visual interpretation of image semantics are not always correct, especially in the existence of low illumination, low contrast, or low image resolution. In comparison, our model incorporates additional discriminative cues (e.g., height, surface normal) to reason about the pixel label assignment, contributing to higher semantic accuracy of image annotations. For the latter, manual annotations often suffer from inaccurate boundaries delimitation in practice, as human annotators can hardly delimit actual segment boundaries for objects with complex shapes such as trees. By contrast, the dense pixelwise connectivity of our model yields accurate label assignments around irregular object boundaries.

A few examples are illustrated in Figure 14. From left to right in each row, Figure 14a–e are snapshots from Google Maps, true orthophotos, manually labeled ground-truth data, corresponding heightmaps and automatically inferred annotations, respectively. Building areas to be noted are highlighted in red. More specifically, figures in the first row show a building which is visible in Google Maps but partly hidden in shadows in the true orthophoto, therefore it was wrongly labeled as ground in manual annotation. However, with the help of the height evidence, our method labeled the building correctly. A similar example is shown in the second row, where the ground was wrongly labeled as building in manual annotation but correctly labeled by our method. The third row presents an example for shape accuracy of buildings, demonstrating that the inferred annotation is more accurate at building boundaries than manual annotation.

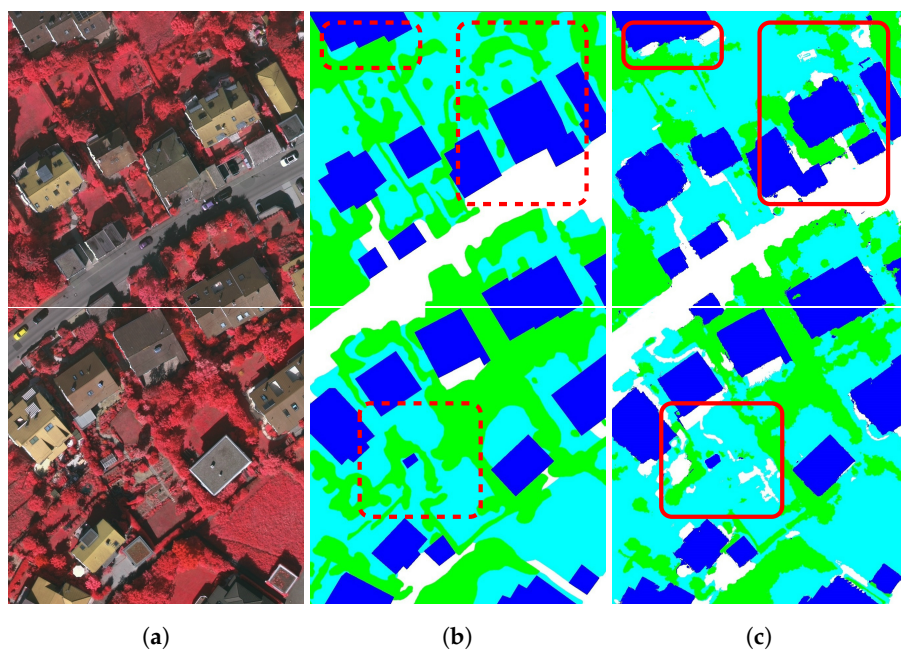
The semantic accuracy of class *Ground* is compared in Figure 15. From left to right in each row, Figure 15a–c are true orthophotos, manually labeled ground-truth data and automatically inferred annotations, respectively. In the two scenes, although the bare ground nearby buildings or through the grassland can be easily recognized on orthophotos, precise annotation is still unpractical due to its irregular shape. Therefore many ground areas are roughly labeled or ignored, as highlighted by the red dashed boxes. By contrast, they are correctly labeled in the automatic annotations, as highlighted by the red solid boxes.

On the other hand, the performance of our method is influenced by the accuracy of heightmaps. Inaccurate height information may lead to wrong labeling, especially when the image content is not distinguishable. For instance, grassland and bushes are labeled respectively as *Low Vegetation* and *Tree* in ground-truth annotations, however, the heightmap extracted from the DSM can not well preserve the small height difference between them and the NDVI evidence is not distinguishable either. This accounts for the wrong annotations for classes *Low Vegetation* and *Tree*, as shown in Figures 14 and 15.





**Figure 14.** Comparison of building accuracy between automatically inferred and manually labeled annotations. (a) snapshot from Google Maps, (b) true orthophotos, (c) manually labeled ground-truth data, (d) corresponding heightmaps, (e) automatically inferred annotations. Building areas to be noted are highlighted in red. Dashed lines indicate wrong labeling and solid lines indicate correct labeling.

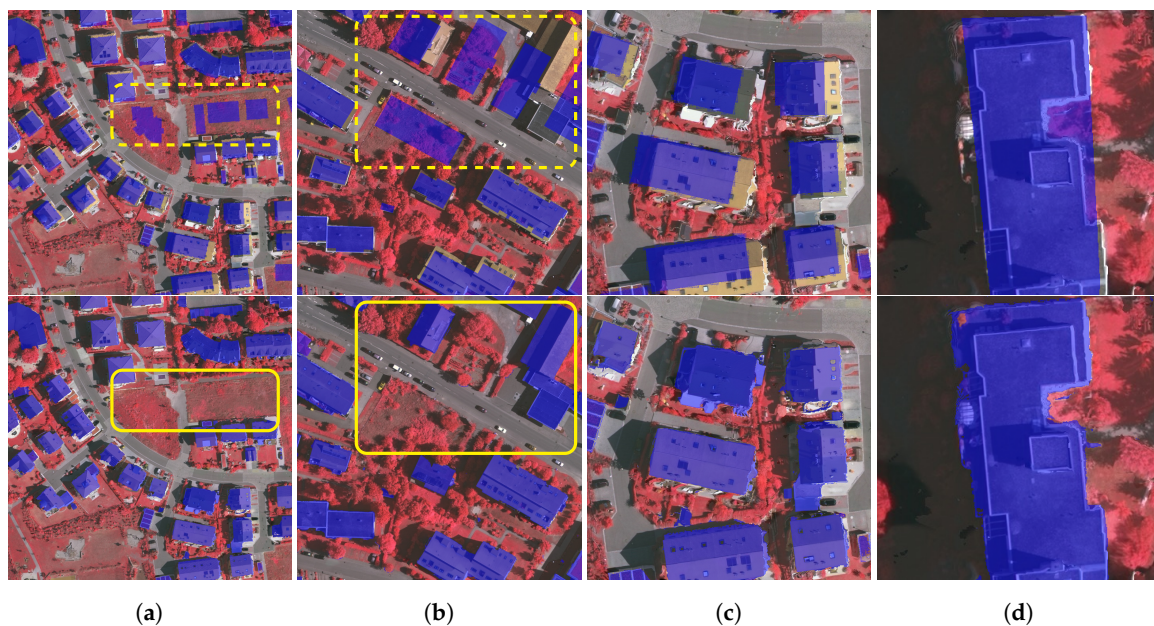


**Figure 15.** Comparison of ground accuracy between automatically inferred and manually labeled annotations. (a) true orthophotos, (b) manually labeled ground-truth data, (c) automatically inferred annotations. Ground areas to be compared are highlighted in red. Dashed lines indicate wrong labeling and solid lines indicate correct labeling.

#### 4.3.2. Comparison with OSM building Footprints

Building annotations are originally derived from the up-to-date OSM data, which has about ten years time offset with the image data. The considerable temporal changes of building result in low semantic

accuracy of the source labeling. Nevertheless, our method is able to generate building annotations with substantially improved accuracy based on additional evidence. Figure 16 presents a comparison between OSM building footprints (top row) and inferred building annotations (bottom row), which are both overlaid on TOP. The dashed rectangles in Figure 16a,b indicate some wrong building labels in OSM footprints caused by temporal changes of buildings, which, in contrast, are correctly labeled in the inferred annotations as highlighted by the solid rectangles. Besides, OSM footprints usually have low position accuracy, especially in rural areas. As compared in Figure 16c, the OSM building footprints have large shift in position while the inferred annotations have apparently higher position accuracy. Third, OSM footprints usually have simplified shapes as illustrated in Figure 16d, but the inferred annotations achieve high shape accuracy for buildings with complex structures.



**Figure 16.** Comparison between OSM building footprints (top row) and inferred building annotations (bottom row), both overlaid on TOP. In (a,b), the dashed rectangles highlight some wrong building labels in OSM footprints and the solid rectangles show the correct labels of inferred annotations. Position accuracy is compared in (c). Shape accuracy is compared in (d).

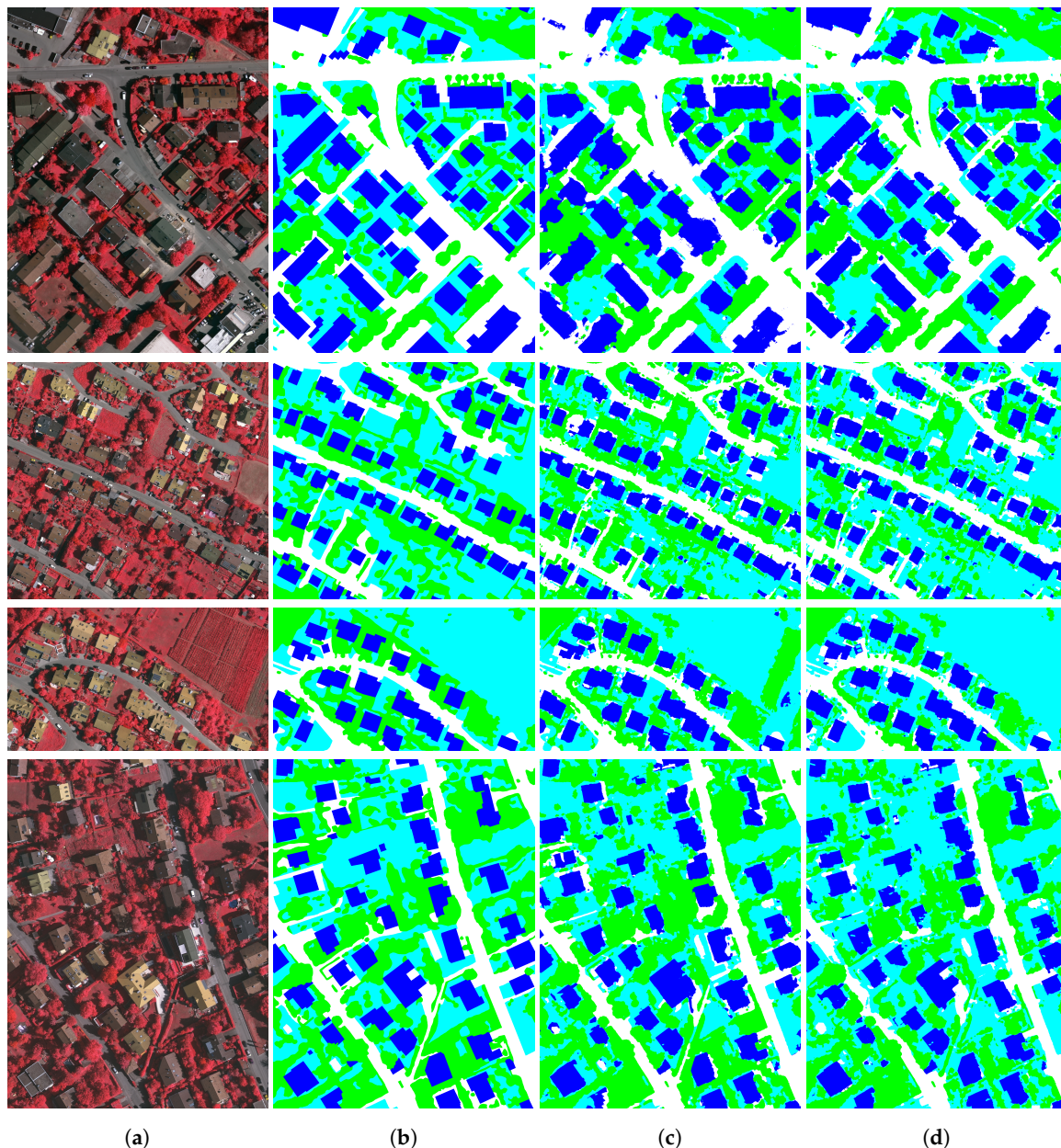
#### 4.4. Training a CNN Using Generated Annotations

In order to validate the utilization of the automatically generated annotations, we leveraged them as training data for image classification using a deep convolutional neural network and compared the performance with the classification using manually labeled training data. Following the training and testing procedures for FCN [27], we selected twelve labeled TOPs (areas: 1, 3, 5, 7, 11, 15, 21, 26, 28, 32, 34 and 37) for training and kept the remaining four labeled TOPs (areas: 13, 17, 23 and 30) for testing. In parallel, we trained another network using the inferred annotations of the 12 tiles, and then compared their performance in testing.

The IoU accuracy of each class is listed in Table 6. In general, classification using inferred annotations achieves comparable overall accuracy as classification using manual annotations. Specifically, classification using inferred annotations achieves higher accuracy for almost all categories except the *Tree* category. Figure 17 depicts full tile predictions using manual and automatic training data, where each row from top to bottom corresponds to tiles No. 30, No. 13, No. 17 and No. 23. From left to right, Figure 17a shows the original true orthophotos, Figure 17b shows corresponding ground-truth, Figure 17c illustrates pixel-wise predictions using manually labeled annotations as training data, Figure 17d depicts pixel-wise predictions using automatically inferred annotations as training data. In order to visualize the distribution of wrong predictions, we also present the corresponding error map in Figure 18. From left to right, Figure 18a shows the original true orthophotos, Figure 18b

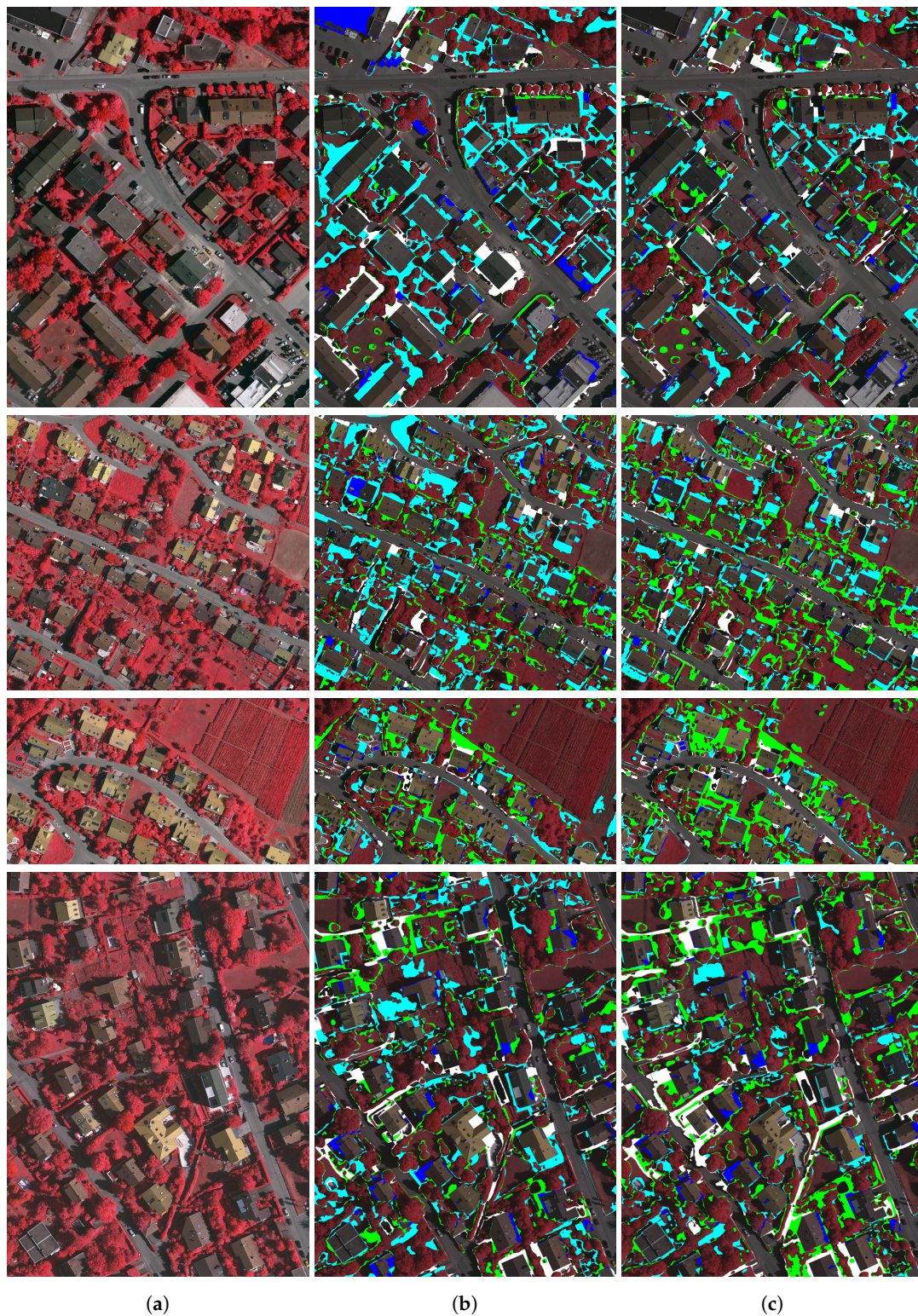


shows the prediction errors using manual annotations as training data and Figure 18c shows the wrong predictions using automatic annotations as training data. Wrong predictions are highlighted in Figure 18b,c and the colors indicate the “true” labels according to the ground truth. For class *Building*, errors are generally caused by shadows and presented as over-segmentations in building areas. For class *Ground* and class *Low Veg*, errors are mainly distributed in shaded areas where the NDVI information cannot well distinguish vegetation and non-vegetation. Errors in class *Tree* generally come from bushes, which have similar height and NDVI with *Low Veg*. In these cases, neither height nor NDVI information is able to well distinguish them. In consequence, many regions of class *Tree* are incorrectly inferred as *Low Veg* or *Ground*.



**Figure 17.** Full tile predictions using manual and automatic training data for Vaihingen dataset. Each row shows from top to bottom: tiles No. 30, No. 13, No. 17, No. 23. (a) true orthophotos, (b) ground truth, (c) predictions using manually labeled annotations as training data, (d) predictions using automatically inferred annotations as training data.





**Figure 18.** Comparison of prediction errors using manual and automatic training data for Vaihingen dataset. Each row shows from top to bottom: tiles No. 30, No. 13, No. 17, No. 23. From left to right, (a) original true orthophotos, (b) prediction errors using manual annotations as training data, (c) prediction errors using automatically generated annotations as training data. Wrong predictions are highlighted using the colors from the ground truth.

Both qualitative and quantitative analysis have demonstrated the effectiveness of leveraging the inferred annotations as ground-truth data for training. In general, image classification using inferred training data is able to achieve comparable accuracy as classification using manual annotations.

**Table 6.** Classification accuracy (IoU) comparison between inferred annotations and manual annotations.

Method	Ground	Building	Low Veg	Tree	Mean
OSM	-	73.70	-	-	
Manual	68.11	84.32	62.08	66.74	70.31
Inferred	70.54	86.78	63.80	59.21	70.08

## 5. Discussion

Compared with other studies on automatic generation of image annotations, we firstly propose the concept of propagating labeled aerial imagery to unlabeled UAV imagery to generate image annotations on a large scale, which can substantially reduce manual labor by utilizing redundant remote sensing data. Although label propagation has been proved to be an effective way to generate image annotations automatically, the state-of-the-art label propagation approaches still face the challenge of inconsistency between source data and target images, especially when they are acquired from different views. Since propagated labels inevitably contain errors, we model labeling uncertainties by introducing additional evidence (e.g., height, NDVI) via the Bayesian inference. In view of the probabilistic nature of our model, we optimize the inferred annotations in a fully connected CRF model defined in image domain. In this context, the proposed method takes advantages of multi-domain information and yield image annotations with both high semantic accuracy and precise boundary partitions.

In view of the low completeness and low accuracy of the initial image annotations, complementary additional information plays an important role for accurate inference. The probabilistic nature of our model allows us to flexibly incorporate different types of evidence, yet appropriate selection and combination of auxiliary information contribute to more accurate and efficient inference. As we integrate additional information by estimating its distribution for each category, it is crucial that each class shows a unique distribution on that attribute or, more broadly, on combinations of attributes. For instance, based on only height values, the ground can be easily distinguished from buildings, but hardly differentiated from the grass as they have similar height distributions; given NDVI values in addition, the three categories can be well discriminated. In this sense, our inference is not restricted to the annotated categories, but can also work even without initial annotations as long as the auxiliary information has distinctive characteristics on each category.

Another merit of the proposed method lies in its ability to preserve precise class boundaries, which is inherited from the characteristics of the CRF model. While manual annotation usually has low accuracy at class boundaries, especially for objects with curvy or complicated shapes such as trees, our method utilize the image contextual information and are inherently sensitive to image gradients, thus the inferred annotations generally exhibit a higher accuracy at class boundaries.

On the other hand, our method also has some limitations. First, the distribution functions of a certain attribute for different objects is empirically defined, its parameters need to be fine-tuned according to the image statistics of the specific dataset. Additionally, our method is sensitive to shadow as it affects the spectral properties of images. In our experiment, errors of the inferred annotations are mostly distributed in shaded areas. Therefore it is advised to perform shadow detection and removal as pre-processing in order to achieve higher semantic accuracy.

## 6. Conclusions

Abundant image annotations are indispensable for training tasks in semantic image segmentation or scene parsing. Traditional image annotation relies on manual labeling, which is quite labor-intensive and unpractical for large-scale tasks. We proposed in this paper a method for automatic image labeling



by label propagation based on a Bayesian-CRF model. In the presence of weak annotations and auxiliary information such as 3D data, our method is able to yield abundant high-quality image annotations in an automatic way. While manual image annotation takes about 30–45 min per frame, our approach can generate image annotations within 1 min. The automatically generated annotations have high semantic accuracy and preserve accurate class boundaries. Besides, the inferred annotations can be used as pseudo ground-truth data for training models. In our experiment, the network trained by inferred image labels achieved comparable classification accuracy as the network trained by manual image annotations, more specifically, the per-class classification accuracy of the networks trained by the manual image annotations and the generated image labels have a difference within  $\pm 5\%$ .

**Author Contributions:** X.Z. contributed to the conceptualization of the work; X.Z. and F.F. proposed the methodology and designed the experiment; F.K. contributed to data acquisition and pre-processing; X.Z. performed the experiments and analyzed the data; X.Z. prepared the original draft and all the authors contributed to reviewing and editing the manuscript. This project is under the supervision of F.F. and the administration of P.R.

**Funding:** This research was funded by the German Academic Exchange Service (DAAD:DLR/DAAD Research Fellowship Nr. 50019750) for Xiangyu Zhuo.

**Acknowledgments:** The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [28]: <http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html>.

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

## References

1. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
2. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [CrossRef]
3. Gaidon, A.; Wang, Q.; Cabon, Y.; Vig, E. Virtual worlds as proxy for multi-object tracking analysis. *arXiv* **2016**, arXiv:1605.06457.
4. Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; Lopez, A.M. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas Valley, NV, USA, 26 June–1 July 2016.
5. Vijayanarasimhan, S.; Grauman, K. Active frame selection for label propagation in videos. In Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 496–509.
6. Sun, C.; Lu, H. Interactive video segmentation via local appearance model. In *IEEE Transactions on Circuits and Systems for Video Technology*; IEEE: Piscataway, NJ, USA, 2017; pp. 1491–1501.
7. Song, J.; Gao, L.; Puscas, M.M.; Nie, F.; Shen, F.; Sebe, N. Joint graph learning and video segmentation via multiple cues and topology calibration. In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 831–840.
8. Badrinarayanan, V.; Budvytis, I.; Cipolla, R. Semi-Supervised Video Segmentation Using Tree Structured Graphical Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2751–2764. [CrossRef] [PubMed]
9. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
10. Badrinarayanan, V.; Galasso, F.; Cipolla, R. Label propagation in video sequences. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3265–3272.
11. Tsai, D.; Flagg, M.; Nakazawa, A.; Rehg, J.M. Motion Coherent Tracking Using Multi-label MRF Optimization. *Int. J. Comput. Vis.* **2012**, *100*, 190–202. [CrossRef]
12. Chen, L.C.; Fidler, S.; Yuille, A.L.; Urtasun, R. Beat the mturkers: Automatic image labeling from weak 3d supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3198–3205.

13. Xiao, J.; Quan, L. Multiple view semantic segmentation for street view images. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 686–693.
14. Namin, S.T.; Najafi, M.; Salzmann, M.; Petersson, L. A Multi-modal Graphical Model for Scene Analysis. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2015; pp. 1006–1013.
15. Xie, J.; Kiefel, M.; Sun, M.T.; Geiger, A. Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas Valley, NV, USA, 26 June–1 July 2016.
16. Mustikovela, S.K.; Yang, M.Y.; Rother, C. Can Ground Truth Label Propagation from Video help Semantic Segmentation? In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
17. Budvytis, I.; Sauer, P.; Roddick, T.; Breen, K.; Cipolla, R. Large Scale Labelled Video Data Augmentation for Semantic Segmentation in Driving Scenarios. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCV Workshop), Venice, Italy, 22–29 October 2017; pp. 230–237.
18. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*; CRC Press: Boca Raton, FL, USA, 2013.
19. Krähenbühl, P.; Koltun, V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. *Adv. Neural Inf. Proc. Syst.* **2011**, 109–117.
20. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), Williamstown, MA, USA, 28 June–1 July 2001.
21. Yang, J.; Yang, M.H. Top-down visual saliency via joint CRF and dictionary learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, 39, 576–588. [[CrossRef](#)] [[PubMed](#)]
22. Kohli, P.; Ladicky, L.; Torr, P.H.S. Robust higher order potentials for enforcing label consistency. *Int. J. Comput. Vis.* **2009**, 82, 302–324. [[CrossRef](#)]
23. Ladický, L.; Russell, C.; Kohli, P.; Torr, P.H.S. Associative hierarchical CRFs for object class image segmentation. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 739–746.
24. Kurz, F.; Rosenbaum, D.; Meynberg, O.; Mattyus, G.; Reinartz, P. Performance of a real-time sensor and processing system on a helicopter. *ISPRS* **2014**, 1, 189–193. [[CrossRef](#)]
25. Zhuo, X.; Koch, T.; Kurz, F.; Fraundorfer, F.; Reinartz, P. Automatic UAV Image Geo-Registration by Matching UAV Images to Georeferenced Image Data. *Remote Sens.* **2017**, 9, 376. [[CrossRef](#)]
26. Mongus, D.; Žalik, B. Parameter-free ground filtering of LiDAR data for automatic DTM generation. *ISPRS J. Photogramm. Remote Sens.* **2012**, 67, 1–12. [[CrossRef](#)]
27. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
28. Cramer, M.J. The DGPF-Test on Digital Airborne Camera Evaluation—Over-view and Test Design. *Photogramm. Fernerkundung Geoinf.* **2010**, 2010, 73–82. [[CrossRef](#)] [[PubMed](#)]
29. Shi, Z.; Siva, P.; Xiang, T. Transfer learning by ranking for weakly supervised object annotation. *arXiv* **2017**, arXiv:1705.00873.2017.
30. Fort, K.; Nazarenko, A.; Rosset, S. Modeling the complexity of manual annotation tasks: A grid of analysis. In Proceedings of the 2012 International Conference on Computational Linguistics, Mumbai, India, 8–15 December 2012; pp. 895–910.

