DIAL: Deep Interactive and Active Learning for Semantic Segmentation in Remote Sensing

Gaston Lenczner, Adrien Chan-Hon-Tong, Bertrand Le Saux, *Senior Member, IEEE*, Nicola Luminari, and Guy Le Besnerais

***Abstract*—In this article, we propose to build up a collaboration between a deep neural network and a human in the loop to swiftly obtain accurate segmentation maps of remote sensing images. In a nutshell, the agent iteratively interacts with the network to correct its initially flawed predictions. Concretely, these interactions are annotations representing the semantic labels. Our methodological contribution is twofold. First, we propose two interactive learning schemes to integrate user inputs into deep neural networks. The first one concatenates the annotations with the other network's inputs. The second one uses the annotations as a sparse ground truth to retrain the network. Second, we propose an active learning (AL) strategy to guide the user toward the most relevant areas to annotate. To this purpose, we compare different state-of-the-art acquisition functions to evaluate the neural network uncertainty such as ConfidNet, entropy, or ODIN. Through experiments on three remote sensing datasets, we show the effectiveness of the proposed methods. Notably, we show that AL based on uncertainty estimation enables to quickly lead the user toward mistakes and that it is thus relevant to guide the user interventions. Code will be open-source and released in this repository.[1]**

***Index Terms*—Active learning (AL), deep learning, earth observation, interactive segmentation, semantic segmentation.**

I. Introduction

A. Context

SEMANTIC segmentation, the task of classifying an image at the pixel level, is extremely important in remote sensing and is addressed with deep neural networks for a variety of applications such as land-cover mapping [1], change detection [2], or farmland monitoring [3]. This task is intrinsically complex, and while deep neural networks can be very effective, they are still prone to failure. Indeed, even on academic benchmarks [4], [5], current state-of-the-art methods often require specific architectures and fine tuning to obtain high performances but still imperfect results. Moreover, it often gets even more tedious on "real-life" datasets due to different factors such as domain adaptation between train and test data inherent to remote sensing data (different weather, geographical areas, types of sensors, cloud shadows, etc.) or the difficulty to have access to large training annotated datasets for every specific business application, even though lots of efforts are made by the community in this direction [6], [7]. Hence, the uncertainty about the quality of the results of neural networks often makes their deployment complicated. Human intervention may then be necessary. Precisely, we are thinking of two scenarios representative of real situations. First, the *Refinement* use case, when the user aims to fix errors within a single dataset, and thus, to improve the performances of the model on the test data. Second, the *domain adaptation* use case, when the user wants to fix on a new dataset, the errors of a model pre-trained on a previous dataset.

A possible way to address these problems comes with interactive learning (IL) [8], [9]. This consists in adding a human in the loop to work in synergy with a learning algorithm to train it, fine tune it, or adapt it to user inputs. Compared to classically supervised algorithms, IL algorithms must also interface smoothly with the human user. This constraint is particularly challenging with deep neural networks due to their typical high number of parameters and long training time. One step further than IL, active learning [10] (AL) searches in pools of unlabeled data, for examples, which are the more able to lead the model to a better classification. These examples, defined as *queries*, are then labeled by the user and incorporated in the training. This thus aims to find the optimal training dataset for the algorithm. To this purpose, AL methods define *acquisition functions* to estimate either the model uncertainty associated to the samples [11] or their representativeness of the dataset [12].

In this article, we explore IL and AL for semantic segmentation. Indeed, as presented in Fig. 1, we propose the deep interactive and active learning (DIAL) framework to interactively refine semantic segmentation maps initially output by a pretrained neural network. First, it relies on two complementary IL schemes to integrate information provided by a user in deep learning algorithms for semantic segmentation. In a nutshell, the first module uses these annotations to modify at test time the inputs of the network pretrained to process them, while the second one uses them for retraining to modify the weights of the network. Second, we integrate AL within our framework and propose to

Manuscript received December 16, 2021; revised February 25, 2022; accepted March 30, 2022. Date of publication April 12, 2022; date of current version May 9, 2022. This work was supported by the company Alteia. (*Corresponding author: Gaston Lenczner*.)

Gaston Lenczner is with the Information Processing and Systems Department (DTIS), ONERA, Université Paris-Saclay, FR-91123 Palaiseau, France, and also with Alteia, FR-31400 Toulouse, France (e-mail: gaston.lenczner@alteia.com).

Adrien Chan-Hon-Tong and Guy Le Besnerais are with the Information Processing and Systems Department (DTIS), ONERA, Université Paris-Saclay, FR-91123 Palaiseau, France (e-mail: adrien.chan_hon_tong@onera.fr; guy.le_besnerais@onera.fr).

Bertrand Le Saux is with ESA/ESRIN Φ-lab, IT-00044 Frascati, Italy (e-mail: bertrand.le.saux@esa.int).

Nicola Luminari is with Alteia, FR-31400 Toulouse, France (e-mail: nicola.luminari@alteia.com).

Digital Object Identifier 10.1109/JSTARS.2022.3166551

[1]https://github.com/alteia-ai/DIAL

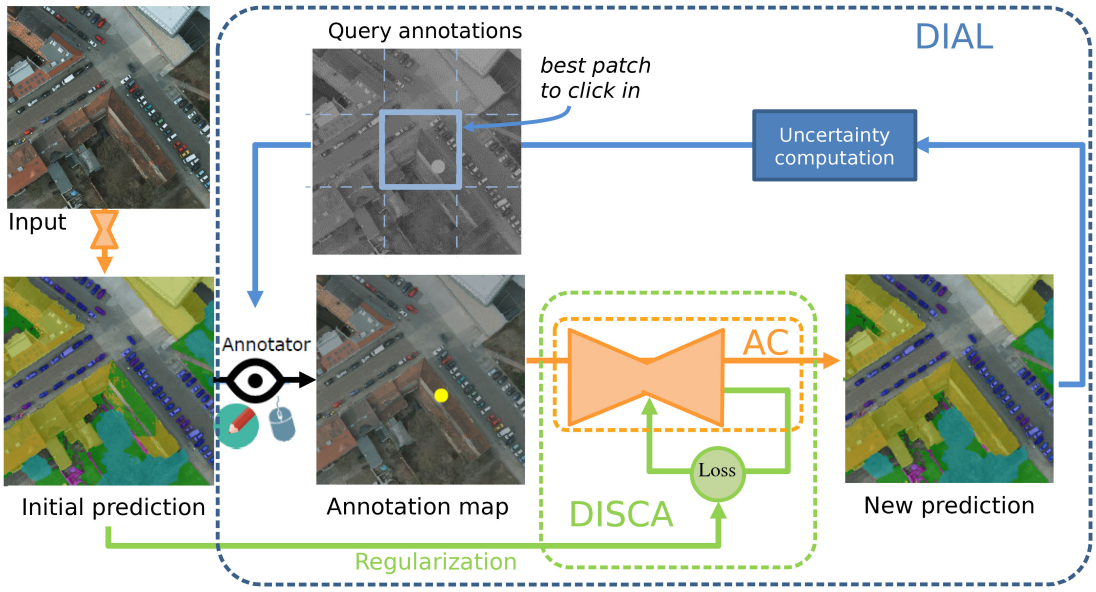


Fig. 1. Visual representation of DIAL encompassing AC, DISCA, and AL. Given a neural network trained to both produce segmentation maps and to use annotations as channels, the framework starts with an initial prediction using the input image without annotations that the annotator can annotate with new points (e.g. to fix errors). Three algorithmic mechanisms cooperate to improve the segmentation map: AC processes jointly image and annotations with the same model without retraining, DISCA additionally retrains the model for better adaptation using the initial prediction as regularization to avoid overfitting on the annotations, and DIAL also proposes most informative patches to speed up the interactions. Best viewed in color. For more details, AC and DISCA are described in Section II-A and the uncertainty-based component in Section II-B.

guide the user interventions toward relevant areas to annotate. This additional guidance relies on different uncertainty measures that we compare with respect to our framework. These measures can be simple-yet-effective such as entropy [13] or come from the current state-of-the-art literature such as ODIN [14] or ConfidNet [15]. We extensively evaluate our framework both in the refinement and in the domain adaptation use cases to well apprehend its potential. We notably show that the first IL module is more suited to correct spatially small mistakes, while the second one is more suited for larger ones and that the active process improves the performance compared to the unguided one.

In summary, the major contributions of this article are as follows:

1) We propose a general framework for interactive multiclass semantic segmentation in remote sensing.
2) We show that AL for area selection allows to speed-up the improvement of the segmentation and reduces the number of required interactions for a given quality.
3) We compare different state-of-the-art methods to estimate the algorithm uncertainty within the DIAL framework sketched in Fig. 1, and show that the model confidence evaluated by ConfidNet and surprisingly the entropy are the most effective, the latter being also the faster one.
4) We show these techniques consistently improve the quality of a segmentation map, with the greater gain for the domain adaption use case where it allows to compensate for the lack of training data in the target domain.

This article extends previous work presented in an international conference [16]. It deepens the IL experiments and combines the interactive and active modules within a single framework.

B. Scenario

We assume the following context for the rest of this article. A user needs to quickly and accurately semantically segment Earth observation images for one of the two aforementioned use cases: the refinement one and the domain adaptation one. This user has also access to another annotated database, which depending on the use case, may or may not belong to the same domain as the targeted images. For the sake of simplicity, the annotated label space must be the same as the targeted one.

We propose to first train a neural network on the annotated database. Then, the user can use this neural network to make predictions on the target images. If the segmentation result is not accurate enough for the user's requirements, they can interact with the network to refine its predictions. These user interactions come in the form of clicked points on the mislabeled areas and represent their corresponding labels, as chosen by the user. Finally, we propose to also guide the user to the most relevant areas of the images using uncertainty estimations relying on different statistical measures.

We have developed a QGIS[2] plugin available with the code to allow potential users to experience the proposed framework. However, to conduct a large-scale evaluation, we have also simulated the user behavior to automatically generate interactions. Hence, in the rest of this article, we refer both to the synthetic operator and to the potential human user as *the agent*.

[2][Online]. Available: http://qgis.osgeo.org

When the agent is simulated to automatically generate the annotations, it samples them in the mistake areas using a comparison between the ground-truth map and the prediction map. It thus necessarily requires a partial access to the ground-truth maps.

To summarize, our framework combines the following three criteria:

1) *Semantic segmentation:* The neural network is able to provide accurate semantic segmentation maps using only the input image without user annotations.
2) *Interactive learning (IL):* The neural network can also refine these segmentation maps using the annotations to efficiently fix its mistakes.
3) *Active learning (AL):* It estimates the neural network uncertainty to guide the user toward queries.

C. Related Work

1) Interactivity in Remote Sensing: Interactive interpretation of remote sensing data has a long history, partially due to the lack of reference data for training in that field. Interactivity has been processed by various techniques to enhance data mining tools with relevance feedback capability : Bayesian modeling of sample distributions was at the core of VisiMine [17], and support vector machines were used in [18]. More recently, boosting and random forests [19] have been the method of choice due to the possibility to train quickly in an incremental manner [9], [20], [21]. Precisely, ALCD [21] trains a random forest on user annotations. With respect to these works, our approach applies deep learning for interactive remote sensing, which is challenging due to the long training time inherent to deep neural networks. Finally, AL has recently been applied to deep learning in remote sensing [22], [23] to interactively update the models. Kellenberger *et al.* [23] address detection of extremely rare objects (e.g., animal detection in aerial images), while Ružička *et al.* [22] deal with rare and varied change detection. On our side, we apply AL to deep learning in the context of segmentation maps refinement.

2) Interactive Segmentation: Interactive segmentation intends to interactively segment an image into foreground and background pixels with user annotations. It was initially addressed using graph-cut-based methods [24] and now mostly by deep neural networks, which take as inputs a concatenation of the RGB image and user annotations [25]. In [26], the authors use the annotations as sparse ground-truth maps to interactively adapt the neural network to a specific object. Multiclass interactive segmentation broadens interactive segmentation to correct multiclass segmentation maps. Agustsson *et al.* [27] proposed a neural network that takes as input a concatenation of the image and the extreme points of each instance in the scene, and then, lets a user correct the proposed multiclass segmentation using scribbles. We do not assume such extreme point map availability as it is extremely costly to acquire in a remote sensing image with potentially many objects.

3) Weakly Supervised Segmentation: When labels are scarce, training usually boils down to learning the most out of the available labels while leveraging unlabeled data to learn a better inner representation as support. To address weakly supervised semantic segmentation in remote sensing, Li *et al.* [28] use image-level labels, while Wang *et al.* [29] focus on domain adaptation using bounding boxes in the target space. Close to semantic segmentation, Daudt *et al.* [30] address change detection in a weakly supervised setting. Semantic segmentation with point supervision was first proposed by What's the Point (WTP) [31], which trains a model from scratch using cross entropy loss on the point labels. Recently and closely related to our work, Hua *et al.* [32] proposed FESTA for weak semantic segmentation in remote sensing. It mainly consists in a regularization to train a neural network from scratch using notably point labels. In this article, we start from a neural network pretrained with full supervision instead of starting from scratch. In our retraining component, we also design a regularization suited to our use cases.

4) Active Learning (AL): AL aims at optimizing the training process of a learning algorithm through an iterative collaboration with a human oracle [10]. It makes the algorithm choose from a pool of unlabeled data, which ones would be the most relevant to improve itself. Then, the oracle provides the asked labels and the algorithm can learn from it. As it defines how to select the data samples to annotate, the acquisition function is the key differentiating component of these methods. These acquisition functions usually rely on an uncertainty, a representativeness or a diversity score computed directly with the model to select the most relevant samples. Uncertainty methods can rely on simple criteria like entropy [13] or disagreement between ensemble models [33] to estimate the model's prediction confidence. As uncertainty-based methods do not aim to be representative of the dataset, they can select very similar examples. To address this issue, representativeness-based methods aim to select the samples in order to form a subset as representative as possible of the entire dataset. Addressing this as a core-set approach, Sener and Savarese [12] solves it like the K-center problem using the L2 distance between the activations of the final fully connected layer of the CNN. Finally, often relying on clustering [34], [35], diversity-based AL aims to select samples that are as diverse (i.e., different) as possible to reduce the redundancy among the selected samples. In the past decade, AL has been deeply explored in remote sensing to train algorithms for animal detection [23], [36], image classification [37], [38], image segmentation [39], and recently for change detection [22]. We borrow from AL techniques to smoothly help the agent to guide our interactive neural network. We focus on uncertainty measures which fit our use case better than representativeness ones since we aim to easily spot wrong predictions and not to increase our training set.

5) Uncertainty in Deep Neural Networks: Uncertainty quantification, or confidence estimation, is a long-standing problem in machine learning and has many applications such as out-of-distribution (OoD) samples detection [14], the decision to trust the model or to defer to a human expertise in fields like healthcare or the detection of new classes in class-incremental learning [40]. Notably, it can also be used in AL to determine which samples should be sent to the oracle for annotation. Many methods to estimate the uncertainty in deep neural networks have

been recently proposed, and they often fall into one of these four categories.

a) Softmax probabilities: The first category of methods uses the probabilities from the softmax output space of the neural networks. Indeed, Hendrycks and Gimpel [41] propose a simple yet strong baseline using the maximum class probability as an uncertainty estimation and apply to outliers detection. However, it is now well-established that softmax probabilities are prone to different issues such as poor calibration [42] and not fit to differentiate in- from out-of-distribution samples [41]. To overcome these issues, Liang *et al.* [14] propose ODIN to detect outliers with a tempered softmax and with adversarial inputs to better distinguish inliers from outliers. Similarly, Lee *et al.* [43] perturb their inputs but instead uses the representation space before the softmax layer and the Mahanabolis distance to do the split.

b) Model ensembling: Due to its intuitive concept and ease of implementation, another popular class of methods estimate the confidence associated to a sample by measuring the disagreement of different models. This model ensembling can either be explicit and use different models [44] or implicit to be less computationally greedy with one stochastic model using dropout [45] (MC Dropout) or batch normalization [22]. However, all these methods inherently require several forward propagation and are thus relatively slow, making them not engaging for interactive interpretation.

c) Auxiliary models: Other recent approaches design an auxiliary model to learn the uncertainty of the downstream model. While [46] mostly focuses on OoD detection, [15] addresses failure prediction and proposes ConfidNet, a neural network to predict if the prediction from the downstream network is accurate or not. These methods do not require to retrain the downstream network and can thus be easily plugged into any prerained architecture. However, they are computationally heavy and require a new training phase for each new task and model. In remote sensing,Rodríguez [47] successfully apply the ConfidNet method for land cover segmentation.

d) Customized loss: Finally, some works design a specific loss to learn the uncertainty directly during training. For instance, Yoo and Kweon [48] train a model to predict the loss associated to a prediction and Moon *et al.* [49] propose a loss that regularizes the class probabilities to better estimate uncertainty. These methods are computationally efficient and model agnostic but require a full training from scratch and cannot be plugged in a pretrained model.

We compare different methods from these categories in this article to optimally guide the agent toward relevant areas to annotate. We focus on the three first categories since an adapted loss is less tailored for our use case as it would require training new models from scratch.

II. DIAL: Deep Interactive and Active Learning

DIAL encompasses different IL modules and an AL module to interactively guide deep neural networks to refine segmentation maps and we now delve in the details of these components, which are also illustrated in Fig. 1.

(a) (b)

Fig. 2. Annotations encoding using distance transform. Best viewed in color. (a) Car (red) and building (yellow) annotations. (b) Distance-transform encoding.

A. IL Components

The three following IL mechanisms form the deep image segmentation with continual adaptation (DISCA) module:

1) Annotations as Channels (AC): The neural network takes as input a concatenation of the RGB image and agent annotations, extending ideas of DIOS [25] to multiclass segmentation. As represented on Fig. 2, the annotations are encoded using distance transform [50] in the annotation channels to better propagate information than with binary clicks. We extensively study different encoding functions in Appendix A. At test time, these annotations are initially provided by the agent in the form of clicked points, then encoded using distance transforms into an N-dimensional tensor, where N is the cardinal of the label space. During the initial supervised training phase, since the neural network needs to learn how to use clicked points as guidance to enhance its initial predictions, we simply provide points randomly sampled from the ground truth to the network. Image-only inputs are also sampled to train segmentation in a standard way and ensure that the network proposes accurate initial segmentation maps. Since the annotations are randomly sampled during training and not specifically in the center of mistake areas, the learned models are able to benefit from clicks independently from their locations. However, from a performance evaluation point of view, placing annotations in large error areas will logically lead to larger metric gains.

2) Retraining on Annotations: Since AC only modifies the network's inputs and not its parameters, the information provided by the annotations does not improve the predictions globally in the image. Inspired by WTP [31], we propose to bypass this locality constraint by retraining the network with a few back-propagation cycles per annotation. Hence, we use the annotations as a sparse ground truth to interactively retrain the network using a cross entropy loss on these annotated pixels. We note $\mathbf{f}$ to represent the neural network parameterized by θ and $\mathbf{x}$ its inputs.

3) Regularization: As only a few pixels are annotated among the millions that usually compose a remote sensing image, the ground-truth maps resulting from the interactions are extremely

sparse. In order to deal with this problem and avoid overfitting, we follow ideas from [26] and [30] by using the initial prediction $\mathbf{p_0} = f(\mathbf{x}, \theta_0)$ for regularization. Precisely we add a $L1$-loss term using the original prediction as reference in order to prevent the model from making a prediction too different from the initial one. Therefore, our loss during the IL process is defined as follows:

$$\mathcal{L}(\mathbf{x}, \mathbf{c}, \mathbf{p_0}; \theta) = \frac{\mathbf{1}_{[\mathbf{c}\neq -\mathbf{1}]}}{\|\mathbf{1}_{[\mathbf{c}\neq -\mathbf{1}]}\|_1} \left\{ -\sum_{i=1}^{N} \mathbf{c}_i \log\left(\mathbf{f}_i(\mathbf{x}; \theta)\right) \right\} + \lambda \|\mathbf{f}(\mathbf{x}; \theta) - \mathbf{p_0}\|_1 \quad (1)$$

where $\mathbf{1}$ represents the indicator function and $\mathbf{c}$ the sparse annotated pixels. In details, $\mathbf{c}$ takes its values in $\{-1, 0, 1\}$. For the pixels annotated as belonging to class i, $\mathbf{c}_i = 1$, and $\mathbf{c}_j = 0$ for all $j \neq i$. For the unannotated pixels, $\mathbf{c}_i = -1$ for all i in $\{1, \ldots, N\}$. $\|\mathbf{1}_{[\mathbf{c}\neq -\mathbf{1}]}\|_1$ weights the loss with respect to the number of annotated pixels. Finally, the positive parameter λ balances the influence of user annotations with respect to the recall toward the initial prediction. Its tuning will be considered in Section III-C.

The two last mechanisms enable the continual learning potential of DISCA and avoid catastrophic forgetting. During the interactive training phase, the AC mechanism is randomly disabled: the annotations are then removed from the inputs. This avoids overfitting on the annotation channels.

Even though AC is part of DISCA, DISCA is deeply different from AC-only due to its retraining component. The corrections are then less localized around the annotations but take more time and this results in quite different outputs. Hence, we see AC-only and DISCA as two distinct interactive schemes that can be easily interchanged and we analyze their respective behaviors throughout Section III.

B. AL Component

Since remote sensing images can be extremely large, DIAL also incorporates an AL strategy to swiftly guide the agent toward queries representing the most meaningful areas of the image to annotate. It is especially adapted to situations where it is difficult to put an *a priori* on the errors of the neural network (i.e., the annotator does not know where to look for errors). With this aim, we compare different state-of-the-art acquisition functions, which estimate the algorithm uncertainty to find the most suited to our usage scenario and interaction setups described as follows:

1) Formalization: To formalize the problem, we note $\mathbf{f}$ to represent the neural network parameterized by θ, $\mathbf{x}$ an input image, $\mathbf{y}$ its associated label map, $\mathbf{a}$ the user annotations, and $\mathbf{g}$ the annotation encoding function. Our goal is then to find the optimal annotations $\mathbf{a}^\star$ minimizing the following problem:

$$\mathbf{a}^\star = \text{argmin}_{\mathbf{a}} \sum_{j \in I} \left(1 - \delta_{\mathbf{y}^j}^{u^j}\right) \quad \text{with } u^j = \text{argmax}_{c \in [[0,N]]} \mathbf{f}_{\theta,c}^j \left(\mathbf{x} \oplus \mathbf{g}(\mathbf{a}, \mathbf{x}, \mathbf{f}_\theta)\right) \quad (2)$$

where $\oplus$ represents the concatenation operation, δ the Kronecker operator, N the cardinal of the label space, and I the pixels set. The problem values range from 0 when all pixels are well classified to $card(I)$ when all pixels are misclassified.

2) Methodology: We propose the following query strategy to benefit from DIAL on a given image. The image is divided into a grid of *N* patches. The patches are annotated consecutively but the order in which they are annotated depends on the uncertainty measure. We have also studied a pixel-based query strategy in Appendix B.

3) Acquisition Functions: We now present the different acquisition functions that we compare to guide the agent.

a) Entropy: We compute the entropy per pixel at the softmax output: $\mathcal{U} = -\sum_c y_c \times \log(f_c(x; \theta))$. As showed by [41], even though the softmax probabilities of a neural network are poorly calibrated, they can still provide a strong baseline to guide the user.

b) MC Dropout: MC Dropout [45] introduces stochasticity in the prediction by enabling dropout regularization at inference time. This allows to obtain an implicit model ensembling. In practice, we add dropout layers in the neural network architecture, and then, make multiple forward passes through the network to create as many softmax vectors. We then compute the variance of these predictions to measure their disagreement and use it as the uncertainty measure.

c) ConfidNet: As proposed by [15], we train a small auxiliary network to learn to estimate the confidence value of the downstream network using its last layers as inputs. It is constituted of one transposed convolutional layer and four 3×3 convolutional layers of, respectively, 32, 120, 64, 32, and 1 output layers. A final sigmoid layer provides the confidence score.

d) ODIN: Following [14] that primarily developed this method for outlier detection, we slightly disturb the image input with an adversarial-like attack aiming to enforce the predicted probabilities of the softmax output toward the predicted classes and add a temperature term in the softmax layer. Then, the adversarial examples are feed forwarded into the network and we use the softmax output maximum class probability as a confidence measure. Formally, we disturb the input with the following perturbation $\mathbf{x} = \mathbf{x} + \varepsilon \Delta_{\mathbf{x}} \mathcal{L}(f_\theta(x), \hat{y})$ where $\mathcal{L}$ represents the cross-entropy loss, $f_\theta(x)$ the predicted probabilities from the softmax output, and $\hat{y}$ the predicted class.

4) Computational Cost: Therefore, these approaches have different inference costs inherent to their underlying structure. Indeed, entropy is virtually cost-free since it computes a simple operation directly on the neural network output. On the contrary, MC Dropout is particularly expensive since it requires computing multiple predictions. Despite the extra prediction, ConfidNet is only slightly more expensive than entropy thanks to the small size of the auxiliary network. Finally, ODIN falls between ConfidNet and MC Dropout due to the creation and inference of the adversarial sample.

III. Experiments

A. Experimental Setup

1) Datasets: We experiment on three semantic segmentation remote sensing datasets: the *INRIA Aerial Image Labelling*

dataset [4] composed of two classes (*buildings* and *not buildings*) and covering more than 800 km^2 in different cities at a 30-cm resolution, the *Aerial Imagery for Roof Segmentation (AIRS) dataset* [51] composed of the same two classes and covering 457 km^2 in New Zealand at a 7.5-cm resolution and the *ISPRS Potsdam dataset* [52] composed of six classes (*impervious surface, buildings, low vegetation, tree, car,* and *clutter*) covering 3 km^2 on Potsdam at a 5-cm resolution. The datasets are divided into a training set and a validation set with a ratio 80%–20%. This allows to synthesize the annotations required to automatically evaluate the framework. During training, the neural network sees 10 000 image slices of size 512×512 randomly sampled from the training set at each epoch for 50 epochs. For evaluation, the images are tiled into patches of size 512×512 with an overlap of size 128 to be processed.

2) Hyperparameters: We use a neural network based on the LinkNet [53] architecture but our approach is agnostic from the neural network backbone.

Except in the annotation encoding study, the annotations are encoded into the neural network channels inputs using distance transform.

For DISCA, during the IL phase, we optimize the weights using ten stochastic gradient descent passes with a learning rate of $2e^{-6}$ and minimize the loss defined in (1) with the regularization parameter λ set to 1.

3) Active Learning (AL) Setup: For ODIN, we set the perturbation parameter ε to $1/255$ and the temperature term to 100.

For MC Dropout, we add a dropout layer between each encoder and decoder block of our architecture, set the dropout rate to 0.1 and compute the variance over five different inferences.

The ConfidNet auxiliary network is trained for ten epochs per dataset with Adam optimizer.

To automatically evaluate the active learning component, we split the test images into 512×512 patches, sample one annotation per patch, and then, make a new prediction on this patch using AC-only and DISCA. With DISCA, we retrain the network sequentially with each patch. We study whether the annotation order can be optimized. The annotations are generated inside the spatially largest mistakes of the patches. We compute the uncertainty globally in the images, and then, compute an uncertainty score per patch by averaging the uncertainty across all the pixels of the patch. We compare the uncertainty-ordered sequences to a randomly drawn one that constitutes the baseline.

B. Performances and Understanding of DIAL Mechanisms

As we can observe on Fig. 3 and on Table I where a 50 clicks budget has been set, both AC and DISCA successfully enhance the outputs initially proposed by the neural network. DISCA reaches better improvement than AC-only on AIRS and ISPRS: in Table I, AC's mean gain with AL is of 2.2%, while DISCA's one is of 2.5%. Visually, this translates into correction of areas as a whole in a single click, like the buildings or the wide plaza of Fig. 3. The slight superiority of not retraining the network with DISCA on INRIA is probably due to a better noise robustness. Indeed, INRIA labeling is based on land register and is thus not as signal compliant as AIRS or ISPRS. The overall superiority of DISCA has to be also moderated by the inference time of the two algorithms. Indeed, as shown in Table II, DISCA is more than $10\times$ slower than AC-only due to its retraining component.

TABLE I
MEAN IOU AFTER 50 ANNOTATED PATCHES WITH RANDOM AND AL (ENTROPY) ORDERS

	Initial	**Rand. patches**		**AL patches**	
	LinkNet	*AC*	*DISCA*	*AC*	*DISCA*
ISPRS	70.7	71.8	71.3	73.1	**73.3**
INRIA	85.4	86.3	86.2	**86.6**	86.4
AIRS	88	88.7	89.4	91.1	**92**

For 50 patches on Figs. 4 and 5, one recovers results from this Table.

TABLE II
MEAN INFERENCE TIME ON A 512×512 PATCH

	Initial	*AC*	*DISCA*
time (s)	0.01	**0.01**	0.11

TABLE III
MEAN TIME FOR PREDICTION WITH UNCERTAINTY COMPUTATION ON 6000×6000 IMAGES

	Random	*Entropy*	*MC Dropout*	*ODIN*	*ConfidNet*
time (s)	9.9	**10.1**	22	15.8	12.6

1) AL With AC: As we can see in Fig. 4, the random order leads to an improvement linear w.r.t. number of processed patches. All AL schemes speed up the gain in performances by targeting the more uncertain areas. This is particularly noticeable on the AIRS dataset where 50 annotations are enough to reach 75% of the final improvement. This behavior is probably due to the dataset itself. Indeed, since it covers a lot of rural areas, many images only contain few buildings and the uncertainty measures then allow to quickly show the areas of interest to the user.

Regarding the different uncertainty measures, ODIN is consistently the worst one. Indeed, it is only slightly better than the random order, and contrary to the other methods, its performance is almost linear on the AIRS dataset. This behavior might be explained by the method original purpose. Indeed, while the other methods aim to estimate the model uncertainty, ODIN aims to detect outliers. Though these tasks are related, it appears here that model errors are not due to this type of issue in the image area. Moreover, Table III shows that ODIN and MCDropout considerably slow the prediction process (resp., by factors 1.5 and 2) compared to entropy (factor 1) and ConfidNet (factor 1.2).

ConfidNet and entropy consistently obtain the best performances, with a slight advantage for the former on AIRS and the domain adaptation use case. However, ConfidNet is also a bit slower and less flexible since it requires to train an additional network for each dataset. Eventually, entropy offers an excellent tradeoff between high accuracy performances and fast computation, as it is only slightly slower than a random pick.

2) AL With DISCA: Since DISCA slightly modifies the neural network parameters, we recompute the entire prediction and uncertainty after each processed patch. Since MC Dropout and

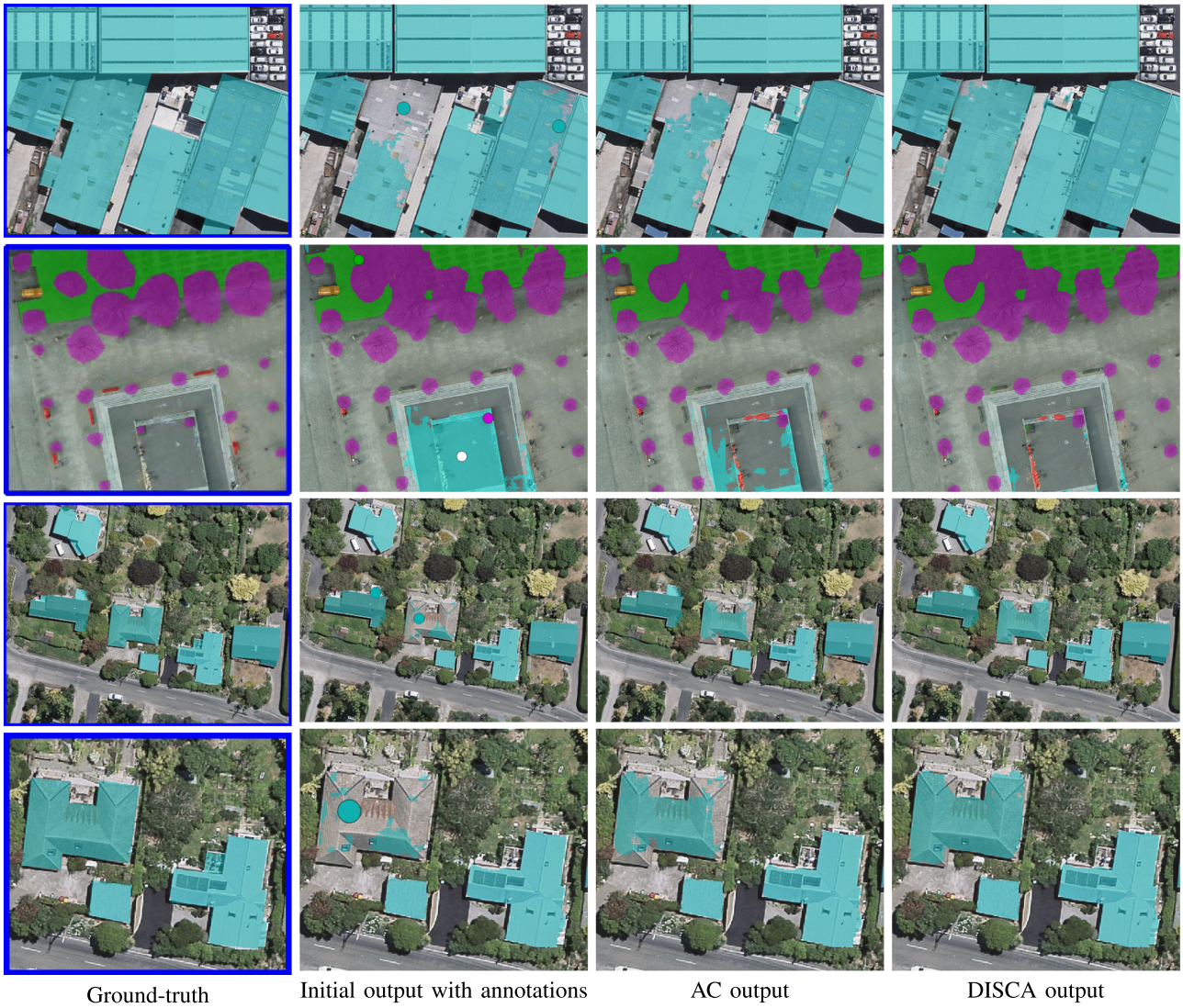


Fig. 3. Visual comparison of the two approaches on examples from AIRS (rows 1, 3, and 4) and ISPRS (row 2). Row 4 is a zoomed version of row 3. In rows 1, 3, and 4, building labels and predictions are in cyan. In row 2, impervious surface labels are transparent and the associated annotations are in white, buildings are in cyan, low vegetation in green, high vegetation in magenta, cars in gold, and clutter in red. In row 2, a wide plaza is initially predicted as "building" and is then corrected as "impervious surface" with also an added "tree" on the right. Best seen in color.

ODIN proved to be relatively slow and less performing with AC, we only compare entropy and ConfidNet in this setup. As we can observe on Fig. 5, results are more complex to interpret than with AC.

On ISPRS, the different methods are all a bit unstable, which is probably explained by the different improvements for the multiple classes of this dataset. However, both uncertainty methods still perform better than the random strategy and the strategy relying on ConfidNet enables a gain up to 5% compared to 4% for the random one. On INRIA, both uncertainty strategies outperform the random one for the first 60 patches but end up being caught up, probably stuck in a local minimum. It is noteworthy that ConfidNet ends up outperforming entropy on these two datasets by a larger margin than with AC. On AIRS and in the domain adaptation situation, the behaviors are similar to the ones obtained with AC, with noticeably higher performances. Indeed, the gain are around 20% and 5% with DISCA, while they were around 10% and 4% with AC, respectively, on the domain adaptation situation and the AIRS dataset.

Hence, these results confirm the benefits of a guidance toward relevant patches relying on uncertainty measures. ConfidNet is on average the best method to this aim. However, the faster, simpler, and only slightly underperforming entropy is a very good alternative for successfully recognizing the most relevant areas to annotate.

C. Ablation Study and Comparison With State-of-the-Art

1) Active Learning (AL): As shown in the previous section, an AL patch order leads to better agent annotations than a random patch order with both AC and DISCA. To this purpose, we compared different state-of-the-art uncertainty-based acquisition functions. We compare it here to a theoretical upper bound of AC and DISCA: the agent generates each click at the center

TABLE IV
PERFORMANCES IN TERMS OF MEAN IOU BEFORE AND AFTER THE INTERACTIVE PROCESSES WITH ONLY TEN ANNOTATIONS PER IMAGE, W.R.T. CORRESPONDING COMPLEXITY

	Initial	**Rand. patches** $\mathcal{O}(n_{\text{annots}} * d^2_{\text{patch}})$		**AL patches** $\mathcal{O}(n_{\text{annots}} * d^2_{\text{patch}})$		**Whole image** $\mathcal{O}(n_{\text{annots}} * d^2_{\text{image}})$	
	LinkNet	*AC*	*DISCA*	*AC*	*DISCA*	*AC*	*DISCA*
ISPRS	70.7	71	68.7	71.8	71.9	71.7	**72.4**
INRIA	85.4	85.5	85.5	86	86	86.4	**86.5**
AIRS	88	88.2	88.8	89.6	**90.7**	89.8	90.2

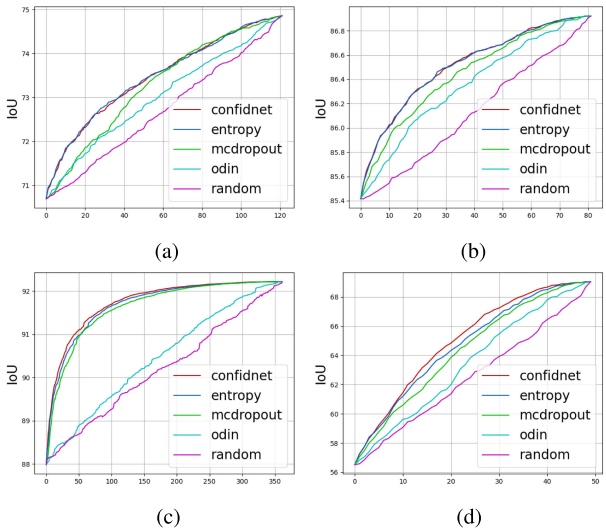


Fig. 4. IoU evolution with respect to the number of annotated patches with AC (one annot. per patch). This compares the different uncertainty measures to select the patch-to-annotate. (a) ISPRS. (b) INRIA. (c) AIRS. (d) AIRS → ISPRS.

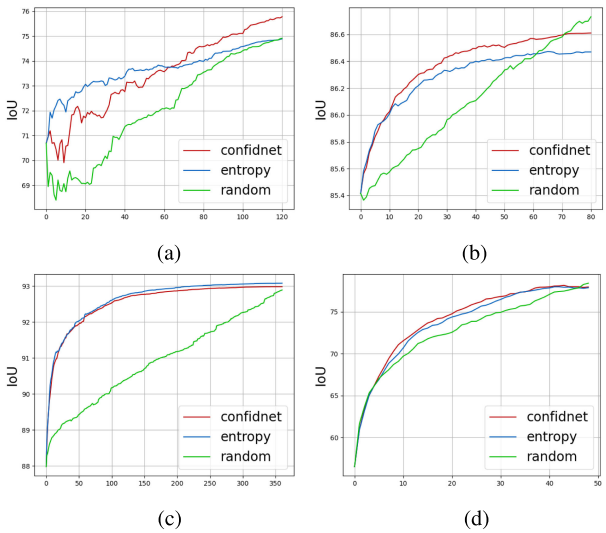


Fig. 5. IoU evolution with respect to the number of annotated patches with DISCA (one annot. per patch). This compares the different uncertainty measures to select the patch-to-annotate. (a) ISPRS. (b) INRIA. (c) AIRS. (d) AIRS → ISPRS.

of the largest spatial error on the whole image, which would be optimal in terms of potential improvement but at the cost of a whole image search. As we can observe in Table IV, this leads,

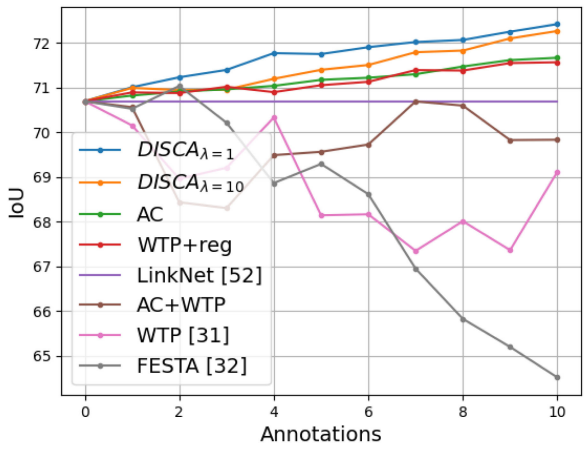


Fig. 6. Ablation study and comparison with the state-of-the-art on ISPRS dataset.

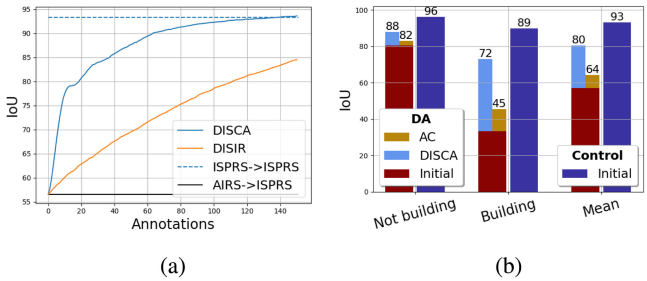


Fig. 7. Mean IoU of AC and DISCA for domain adaptation (AIRS → ISPRS). The fully supervised model (ISPRS → ISPRS) is outperformed in (a) by DISCA after 130 annotations. (a) Mean IoU w.r.t the number of annotations. (b) IoU after 20 annotations.

with ten annotations with AC/DISCA, to an average 1.3/1.7% improvement over the three datasets against a 1.1/1.5% improvement with the AL strategy. However, this slight superiority is mitigated by the complexity to find the annotations. Indeed, in the whole image case, the agent has to browse through 3.6×10^7 pixels for each click in a 6000×6000 image (complexity: $\mathcal{O}(n_{\text{annots}} * d^2_{\text{image}})$), while, in the patch case, it has to browse through 2.6×10^5 pixels in a 512×512 patch (complexity: $\mathcal{O}(n_{\text{annots}} * d^2_{\text{patch}})$). Hence, it is 100 times more costly to find the annotation in an entire remote sensing image than in a patch.

2) AC and DISCA: To understand the influence of the aspects of the DISCA algorithm, we analyze separately its different components, all of them coming from state of the art works. AC (ours) adds input layers for annotations [25], randomly pretrained on the ground truth. WTP [31] retrains the model based on a few annotations. DISCA (ours) sums up AC and

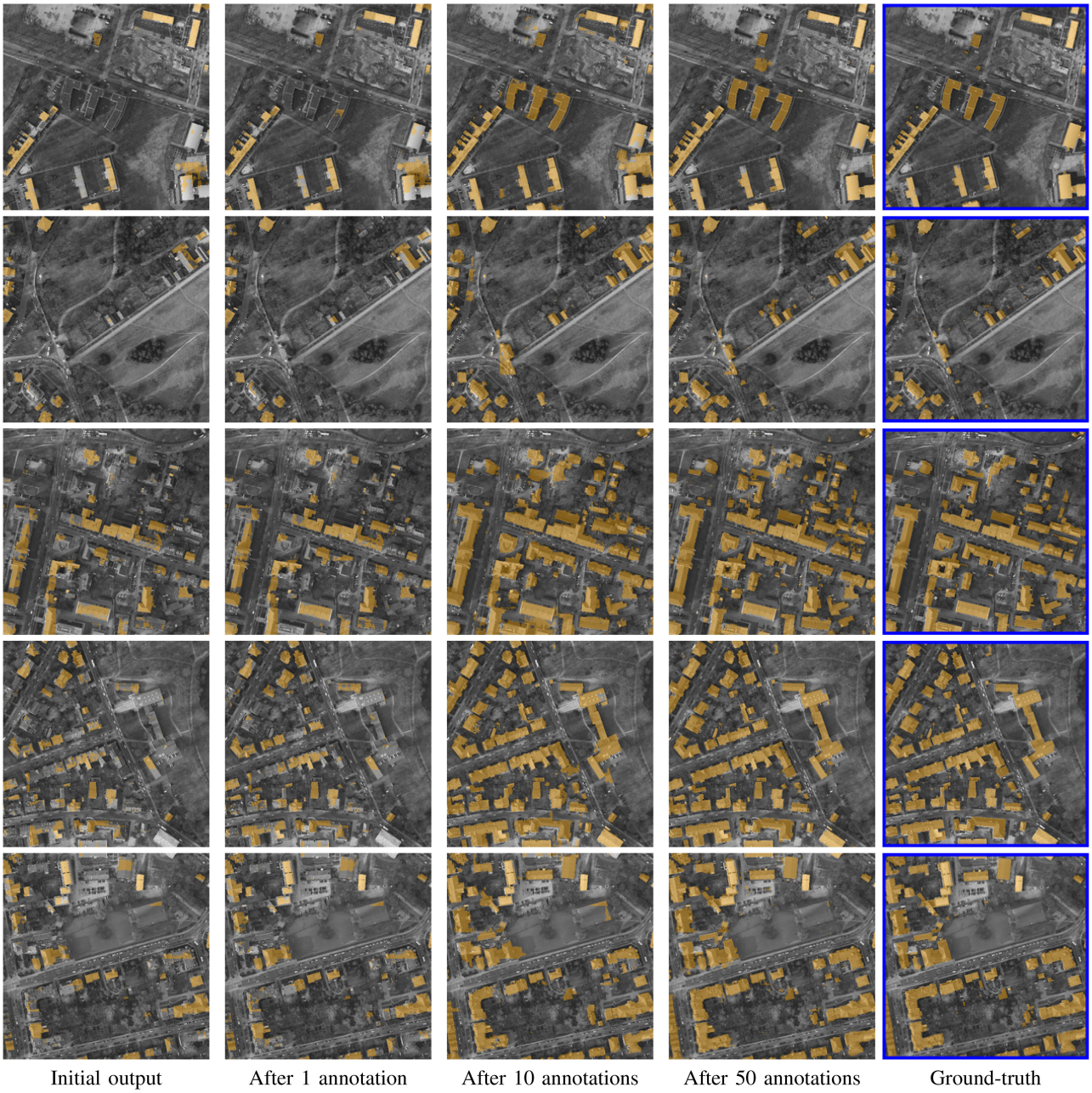


Fig. 8. Domain adaptation (AIRS → ISPRS) visual examples

WTP with regularization with respect to the initial prediction. We also test AC combined with WTP, and WTP combined with regularization. To study the importance of the regularization parameter λ, we test various values $\text{DISCA}_{\lambda=1}$ and $\text{DISCA}_{\lambda=10}$. Finally, we also compare our models to FESTA [32], which trains a neural network on point annotations (as WTP) with a different regularization.

As shown on Fig. 6, AC and WTP+reg obtain IoU gains around 1% for ten clicks and are beaten by the various flavors of DISCA, which almost doubles the gain. This means that the interactive retraining process could be effectively applied to any classically trained neural network but needs to be combined with the AC process to fully exploit the annotations. Moreover, we observe that the regularization is extremely important in DISCA as its absence leads to worse results (AC+WTP curve) than the initial ones (LinkNet curve). A too high λ also decreases the benefits brought by DISCA because it then prevents the algorithm to optimally exploit the annotations. Finally, in this framework of incremental learning, WTP [31] and FESTA [32] also lead to worse results than the initial ones, as emphasized in Table V. These methods were originally designed to train the neural networks from scratch on point annotations. Hence, it explains why they are not optimal in a refinement scenario since they take into account different constraints.

We also compare our approaches with the recent ALCD method [21] also deployed in the field of remote sensing for

TABLE V
COMPARISON ON ISPRS DATASET AFTER TEN ANNOTATIONS

	mIoU
WTP [31]	69.1
FESTA [32]	64.5
AC (ours)	71.6
DISCA (ours)	**72.4**

cloud segmentation in low resolution (60 m/pixel) images. To adapt it to our use case, we run ALCD in a fine-tuning setting on the ISPRS dataset. In practice, we initially pretrain the ALCD random forest on 100 000 samples per image from the training set, and then, adapt the classifier with the same number of annotations as AC and DISCA. However, it leads to very poor performances both before (30% IoU) and after fine tuning (30.5% IoU) compared to AC/DISCA results presented previously. While the absolute results might be due to differences of peculiar implementations of random forest and neural network, the ALCD gain is only +0.5%, which is two times less than AC and three times less than DISCA.

D. Domain Adaptation Use Case

1) Performances: The objective in this domain adaptation use case is to detect the buildings on the eight images of the ISPRS validation set. To this purpose, we compare a neural network trained on AIRS under AC and DISCA settings to a control one trained on the ISPRS training set. The ISPRS images are down-sampled using bilinear interpolation to the AIRS resolution. The neural network's weights are reinitialized between each image. Fig. 7 shows that a network weakly supervised with DISCA beats AC by a large margin in this scenario. Besides, it can quickly reach high performances (more than 80% IoU within 20 annotations) and even outperform a fully supervised one with a sufficient amount of annotations. This is visually confirmed on Fig. 8. Indeed, ten annotations enable the network to well understand the new domain images, and thus, propose decent segmentation maps. More annotations correct most of the remaining mistakes.

2) Sequential Learning: Moreover, we analyze the generalization of DISCA through a sequence of images in the same domain adaptation scenario. This means that we do not reinitialize the neural network weights between each image. We refer to this set up as sequential learning, and we learn two insights from it on Fig. 9. First, DISCA does not suffer from catastrophic forgetting here as the algorithm does not diverge even on the last seen images. Second, sequential learning greatly improves the initial performances directly after the first image. Indeed, the initial IoU then approximately increases by 20%. However, after few annotations, the sequential learning benefits vanish and the performances become similar to the nonsequential setup.

E. Discussion: When to Choose AC-Only or DISCA?

To better apprehend the difference between the two methods, we sample 10 000 512×512 crops from each dataset. Then, given one annotation, we compare the difference between AC and DISCA based on two parameters: the spatial size of the corrected mistake and the initial accuracy of the network on the patch. Precisely, the spatial size of the corrected mistake is the size of the error polygon in which the annotation is sampled. Similarly to the initial accuracy, it is obtained with a comparison between the initial predicted map and the ground-truth map. It is intuitively obvious that both AC and DISCA are correlated to these parameters since, if the mistake to correct is small, the overall IoU gain will be smaller than with a larger mistake to correct. However, we think that this comparison can bring valuable insights to choose the appropriate method depending on the situation.

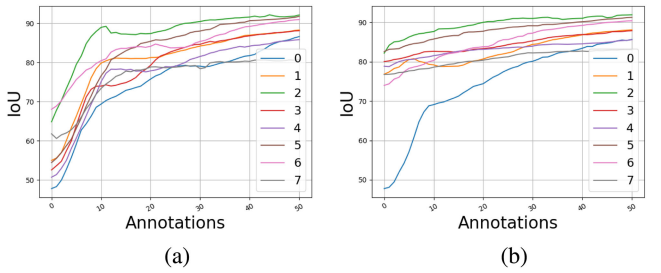


Fig. 9. Sequential learning study with DISCA in a transfer scenario. The legend corresponds to the order in which the algorithm processes the images. (a) Weights reinitialized between each image. (b) Weights updated between each image.

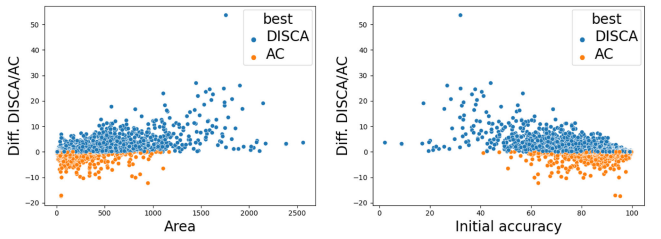


Fig. 10. Comparison of AC-only and DISCA (IoU) with respect to the spatial size of the corrected mistake and the initial accuracy. Legend "best" designates the best method for the given sample.

Fig. 10 compares the two methods with respect to these two criteria. First, both methods seem to work well and can outperform the other one when the mistake area is small and the initial performance is high. We thus recommend to use AC in these situations. Indeed, the locality of AC is no longer a constraint since the error is strongly spatially contained and the relatively long retraining time inherent to DISCA makes it less suitable here. Second, when the initial accuracy is low or the area to correct large, DISCA now clearly tends to perform better than AC, and we thus believe that it should be favored in these situations. Indeed, its spatial globality resulting from its retraining can be fully expressed to correct large mistakes. This outcome shows that DISCA is more relevant to correct deeply flawed segmentation maps than AC.

IV. CONCLUSION

In this article, we have presented DIAL, a framework to interactively enhance segmentation maps initially proposed by

a neural network. Its core concept relies on interactions between a deep neural network and an agent under the form of clicked annotations.

First, we have proposed an IL framework that builds on complementary mechanisms. First, AC modifies the neural network inputs. This approach is fast and local since it does not modify the weights of network. Second, an on-the-fly retraining uses the annotations as a a sparse ground truth. Finally, a regularization term based on the initial prediction is crucial to complement the cross-entropy loss during retraining and avoids catastrophic forgetting. Since this modifies the weights of the network, the full framework is slower but improves the segmentation maps at a larger scale.

Finally, we have integrated AL within our framework to guide the agent interventions toward relevant patch queries. To this purpose, we have compared different state-of-the-art acquisition functions to estimate the neural network uncertainty to finally conclude that entropy is the most suited one thanks to its simplicity and efficiency. Hence, AL speeds up the use of our interactive segmentation algorithms and is particularly relevant to face budget constraints.

In the future, we intend to apply DIAL in a class-incremental scenario to make it easily adaptive to new tasks and use cases. This, along with the promising results shown in domain adaptation, will provide Earth observation scientists and companies with a powerful to reuse, transfer, and enhance deep learning models.

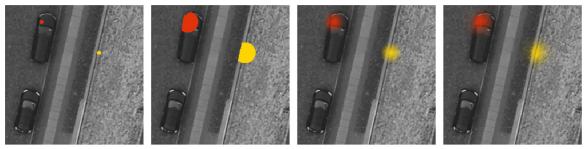
APPENDIX A
HOW TO OPTIMIZE THE ENCODING ?

We investigate here the annotations encoding to analyze its influence on the AC mechanism.

There are many possibilities to encode the annotations in their dedicated channels and they all provide different spatial information. The size of the encoding is the most obvious issue: if it is too small, it might not provide enough information to efficiently fix the initial segmentation but a coarser encoding might provide erroneous information. A popular context-free tradeoff used in most interactive segmentation works [25], [54] is to encode the annotations with Euclidean distance transforms to dilute spatial information. However, due to its context independence, this encoding might be suboptimal. Ideally, the perfect encoding would be the original ground-truth map but it is obviously impossible to get. Based on this insight, we study here how to best approximate this ground truth given the available data: the input image, the annotations, and the trained neural network. We define the following two possible context use besides the no-context one:

1) using the input image;
2) using the initial prediction.

As encoding baselines, we use small binary (bin.) disks of 1.5-pixels radius and distance transform (DT) applied on 10-pixels radius disks. We build on this DT encoding for the context encodings. To use the input image, we rely on guided filtering (GF) [55] in order to preserve the edges in the encoding. To use the initial prediction, we encode the annotations using their connected pixels in the prediction map (C-PM). To estimate the superior boundary theoretically reachable with an encoding from the ground truth, we also encode the annotations using their connected pixels in the ground-truth map (C-GT). These different methods to encode the annotations are represented in Fig. 11.

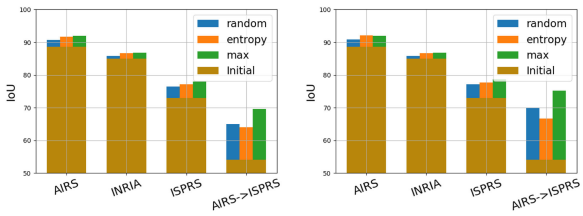


(a) (b) (c) (d)

Fig. 11. Different annotations encodings depending on context uses. Best viewed in color. (a) Car (red) and building (yellow) annotations. (b) Ground-truth connectivity encoding. (c) Distance-transform encoding. (d) Guided filter encoding.

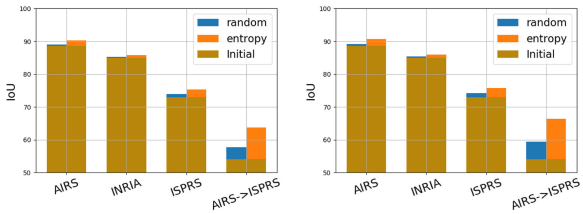
TABLE A1
IOU ON ISPRS AFTER 120 ANNOTATIONS WITH AC DEPENDING ON THE ENCODING

	Bin.	*DT*	*C-PM*	*C-GT* (sup)	*GF*
Initial	70.7	70.7	70.7	70.7	70.8
After	76.4	76.6	76.5	**76.7**	76.7
Gain	5.7	5.9	5.8	**6**	5.9



(a) (b)

Fig. 12. Annotations sampled using uncertainty and error knowledge (smart corrector agent). (a) AC-only. (b) DISCA.



(a) (b)

Fig. 13. Annotations sampled with uncertainty but without error knowledge. (a) AC-only. (b) DISCA.

However, as shown in Table A1, the different encoding strategies seem to provide similar information to the network as the gains are in the same order of magnitude. Indeed, they all increase the IoU of around 6% for 120 annotations on the ISPRS images, even though the binary encoding is slightly lower and confirms the usefulness of DT encoding. The GF encoding obtains the same gain as the DT one, C-PM is lower of 0.1% and even C-GM is only better of 0.1%.

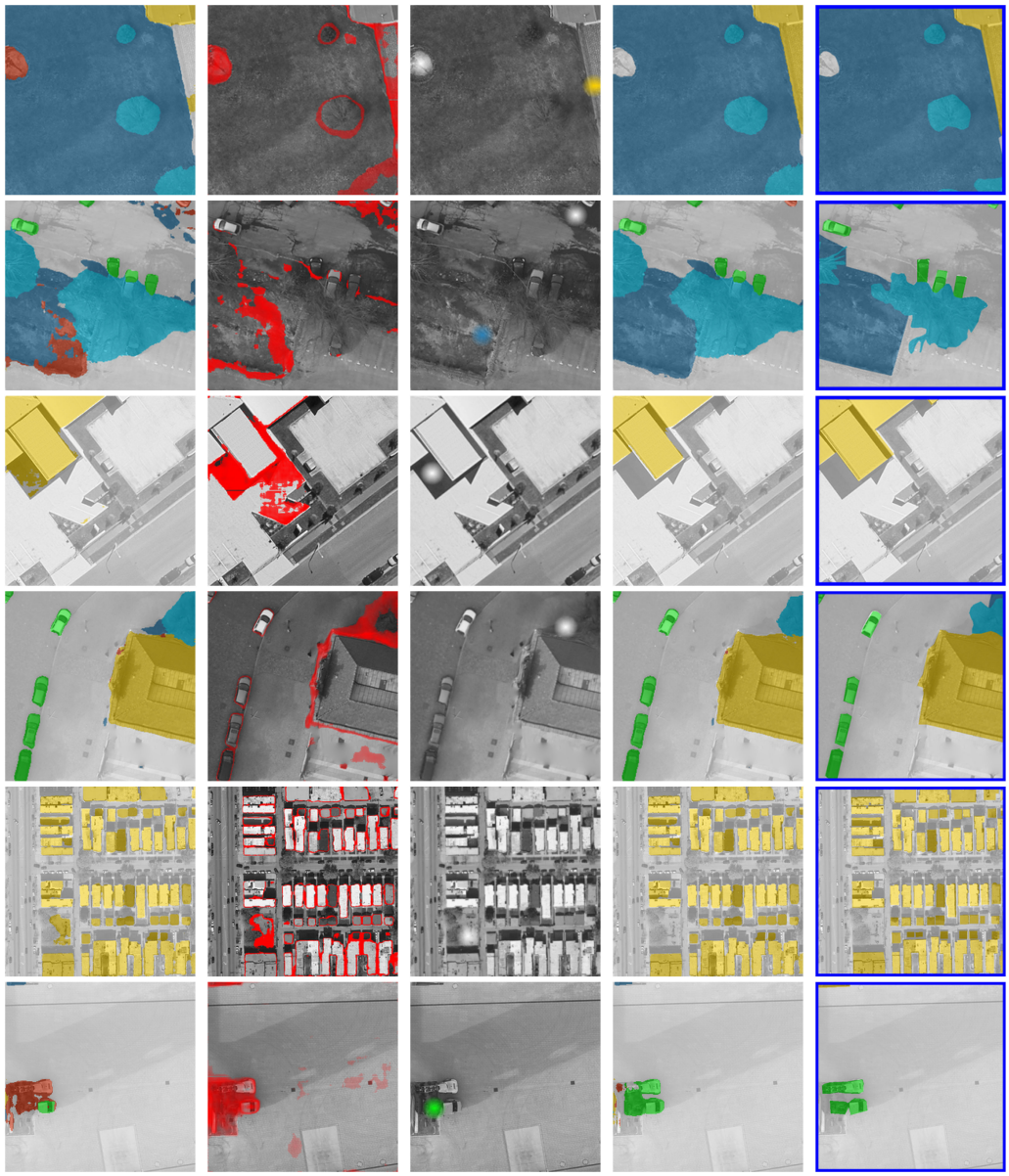


Fig. 14. Initial output corrected with annotations relying on entropy. On the "*Entropy*" column, the areas with an entropy higher than the ninth quantile over the image are highlighted in red. On the "*Annotation(s)*" column, the color of the annotations represents their labels w.r.t. the associated ground-truth maps.

These insignificant differences show that the network does not need any contextual guidance to learn nearly optimal information from the annotations using a simple and intuitive encoding such as distance transform.

APPENDIX B
AL FOR PIXEL-BASED GUIDANCE

A. Setup and Objectives

To investigate local guidance, we sample 10 000 512×512 crops from each dataset, make an initial prediction, generate one annotation per sample, and then, do a new prediction with ac-only and DISCA. Our acquisition function here is entropy. We test two conjectures. First, we want to determine whether highly uncertain pixels among the misclassified ones can lead toward particularly meaningful annotations. Second, we want to figure out whether the uncertainty measurements can help the agent to spot errors at a pixel level.

B. Uncertainty for Optimal Annotations

We make the hypothesis that an agent always clicks on a wrongly segmented area, or in other words that he is able to spot the mistakes and correct them. To look for optimal annotations, we compare the following annotations sampling strategies.

1) We sample the annotation randomly in the wrongly segmented area (*random*).
2) Like in the other experiments, we sample the annotation in the middle of the spatially largest wrongly segmented area (*max*).
3) We threshold the uncertainty map at the ninth quantile to keep only the highest uncertainty values. We then

sample the annotation in the intersection of the wrongly segmented area and this thresholded uncertainty map.

As shown on Fig. 12, the uncertainty-based annotations lead to corrections of the same magnitude than the random ones on average. Moreover, these uncertainty-based annotations clearly do not provide more information to the model than the ones based on *max*. Indeed, the gains of *max* annotations with AC-only are around 6.4% IoU against, respectively, 4.5% and 4.7% for the random and uncertainty-based ones.

This corroborates the correlation between the gain and the size of the corrected area previously exhibited and shows that uncertainty does not lead toward more meaningful annotations than the ones contained inside large mistakes.

C. Uncertainty to Spot Mistakes

In order to evaluate if the uncertainty measures can help to spot mistakes at the pixel level, we compare annotations sampled randomly and on the basis of uncertainty measures *without* ground-truth prior knowledge. In other words, we do not coerce the annotations to be sampled in mistake areas.

Fig. 13 shows that the uncertainty-based annotations lead to better improvements (3.7% IoU with AC-only, 4.5% with DISCA) than the random ones (1.3% IoU with AC-only, 1.9% with DISCA) on average. We can visually confirm these insights on Fig. 14 where uncertainty measures tend to highlight wrongly predicted areas. Besides, the highlighted areas that are initially correctly predicted tend to be legitimately questionable such as object contours or road surfaces looking like buildings (third row).

References

[1] H. Costa, G. M. Foody, and D. S. Boyd, "Supervised methods of image segmentation accuracy assessment in land cover mapping," in *Remote Sensing of Environment*, vol. 205. Amsterdam, The Netherlands: Elsevier, 2018, pp. 338–351.

[2] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[3] M. T. Chiu *et al.*, "Agriculture-vision: A large aerial image database for agricultural pattern analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2828–2838.

[4] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The INRIA aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

[5] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[6] J. Castillo-Navarro, B. Le Saux, A. Boulch, N. Audebert, and S. Lefèvre, "Semi-supervised semantic segmentation in Earth observation: The MiniFrance suite, dataset analysis and multi-task network study," *Mach. Learn.*, pp. 1–36, 2021.

[7] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, 2019, pp. 42–55.

[8] M. Schroder, H. Rehrauer, K. Seidel, and M. Datcu, "Interactive learning and probabilistic retrieval in remote sensing image archives," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 5, pp. 2288–2298, Sep. 2000.

[9] B. Le Saux and M. Sanfourche, "Rapid semantic mapping: Learn environment classifiers on the fly," in *Proc. Int. Conf. Intell. Robots Syst.*, 2013, pp. 3725–3730.

[10] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. TR1648, 2009.

[11] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1183–1192.

[12] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *Int. Conf. Learn. Represent.*, 2018.

[13] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[14] S. Liang, Y. Li, and R. Srikant, "Principled detection of out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 655–662.

[15] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez, "Addressing failure prediction by learning model confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 2902–2913.

[16] G. Lenczner, A. Chan-Hon-Tong, N. Luminari, B. Le Saux, and G. Le Besnerais, "Interactive learning for semantic segmentation in Earth observation," in *Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discov. Databases Workshop*, 2020.

[17] S. Aksoy, K. Koperski, C. Tusk, and G. Marchisio, "Interactive training of advanced classifiers for mining remote sensing image archives," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Databases*, 2004, pp. 773–782.

[18] M. Ferecatu and N. Boujemaa, "Interactive remote-sensing image retrieval using active relevance feedback," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 818–826, Apr. 2007.

[19] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[20] J. A. dos Santos, P. Gosselin, S. Philipp-Foliguet, R. d. S. Torres, and A. X. Falcão, "Interactive multiscale classification of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 4, pp. 2020–2034, Aug. 2013.

[21] L. Baetens, C. Desjardins, and O. Hagolle, "Validation of Copernicus sentinel-2 cloud masks obtained from MAJA, Sen2Cor, and FMask processors using reference cloud masks generated with a supervised active learning procedure," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 433.

[22] V. Ružička, S. D'Aronco, J. D. Wegner, and K. Schindler, "Deep active learning in remote sensing for data efficient change detection," in *Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discov. Databases Workshop*, 2020.

[23] B. Kellenberger, D. Marcos, S. Lobry, and D. Tuia, "Half a percent of labels is enough: Efficient animal detection in UAV imagery using deep CNNs and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9524–9533, Dec. 2019.

[24] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[25] N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang, "Deep interactive object selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 373–381.

[26] T. Kontogianni, M. Gygli, J. Uijlings, and V. Ferrari, "Continuous adaptation for interactive object segmentation by learning from corrections," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 579–596.

[27] E. Agustsson, J. R. Uijlings, and V. Ferrari, "Interactive full image segmentation by considering all regions jointly," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11622–11631.

[28] Z. Li, X. Zhang, P. Xiao, and Z. Zheng, "On the effectiveness of weakly supervised semantic segmentation for building extraction from high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3266–3281, Mar. 2021, doi: 10.1109/JSTARS.2021.3063788.

[29] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4376–4386, Sep. 2019.

[30] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Guided anisotropic diffusion and iterative learning for weakly supervised change detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2019.

[31] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 549–565.

[32] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuia, "Semantic segmentation of remote sensing images with sparse annotations," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Jan. 2021, doi: 10.1109/LGRS.2021.3051053.

[33] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, Oct. 1990.

[34] H. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proc. Int. Conf. Mach. Learn.*, 2004, Art. no. 79.

[35] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1014–1031, Mar. 2011.
[36] M. Laroze, R. Dambreville, C. Friguet, E. Kijak, and S. Lefèvre, "Active learning to assist annotation of aerial images in environmental surveys," in *Proc. Int. Conf. Content-Based Multimedia Indexing*, 2018, pp. 1–6.
[37] D. Tuia, F. Ratle, F. Pacifici, M. Kanevski, and W. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.
[38] L. Bruzzone and C. Persello, "Active learning for classification of remote sensing images," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2009, pp. 689–693.
[39] J. Guo, X. Zhou, J. Li, A. Plaza, and S. Prasad, "Superpixel-based active learning and online feature importance learning for hyperspectral image analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 1, pp. 347–359, Jan. 2017.
[40] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2001–2010.
[41] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Representations*, 2017.
[42] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.
[43] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 7167–7177.
[44] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9368–9377.
[45] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.
[46] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," 2018, *arXiv:1802.04865*.
[47] C. G. Rodríguez, J. Vitrià, and O. Mora, "Uncertainty-based human-in-the-loop deep learning for land cover segmentation," *Remote Sens.*, vol. 12 no. 22, 2020, Art. no. 3836.
[48] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 93–102.
[49] J. Moon, J. Kim, Y. Shin, and S. Hwang, "Confidence-aware learning for deep neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7034–7044.
[50] N. Audebert, A. Boulch, B. Le Saux, and S. Lefèvre, "Distance transform regression for spatially-aware deep semantic segmentation," *Comput. Vis. Image Understanding*, vol. 189, 2019, Art. no. 102809.
[51] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. Waslander, "Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," *ISPRS J. Photogramm. Rem. Sens.*, vol. 147, pp. 42–55, 2019.
[52] F. Rottensteiner *et al.*, "The ISPRS benchmark on urban object classification and 3D building reconstruction," in *Proc. ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 1, no. 1, pp. 293–298, 2012.
[53] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. Conf. Vis. Commun. Image Process.*, 2017, pp. 1–4.
[54] J. Liew, Y. Wei, W. Xiong, S.-H. Ong, and J. Feng, "Regional interactive image segmentation networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2746–2754.
[55] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.

Gaston Lenczner received the M.Sc. degree in statistics and data science from Sorbonne Université, Paris, France, in 2018. Since 2019, he has been working toward the CIFRE Ph.D. degree with the Université Paris-Saclay, in collaboration with Alteia and ONERA, Palaiseau, France.

His research interests include interactions between deep learning algorithms and their users for scene understanding and Earth observation applications.

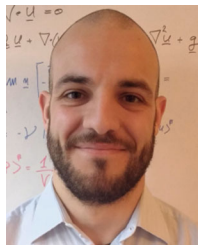
Adrien Chan-Hon-Tong received the graduate degree in machine learning from Ecole Polytechnique, Palaiseau, France, in 2011, and the Ph.D. degree from Sorbonne Universite (in collaboration with CEA), Paris, France, in 2014.

Since 2014, he has been working with ONERA, Palaiseau. From 2014 to 2018, he mainly worked on real-time aerial detection (for defense purpose). This made him to start thinking about safety issues raised by machine learning on critical platform (autonomous driving and lethal autonomous weapon systems). His current research interests include Trustworthy AI: data poisoning, adversarial attack, performance on poor condition, and certification process.

Bertrand Le Saux (Senior Member, IEEE) received the Ms.Eng. and M.Sc. degrees in electrical engineering from Institut National Polytechnique, Grenoble, France, in 1999, the Ph.D. degree in computer science from the University of Versailles/Inria, Versailles, France, in 2003, and the Dr. Habilitation degree in physics from the University of Paris-Saclay, Saclay, France, in 2019.

He is a Senior Scientist with the European Space Agency/European Space Research Institute Φ-lab, Frascati, Italy. His research interest include visual understanding of the environment by data-driven techniques including artificial Intelligence and (Quantum) machine Learning, and tackling practical problems that arise in Earth observation, to bring solutions to current environment and population challenges.

Dr. Le Saux is an Associate Editor of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He was the Co-Chair in 2015–2017 and the Chair in 2017–2019 for IEEE GRSS Technical Committee on Image Analysis and Data Fusion.



Nicola Luminari received the Ph.D. degree in fluid dynamic from Institut national Polytechnique de Toulouse, Toulouse, France, in February 2018, with the thesis on "modeling and simulation of flows over and through fibrous porous media" involved mathematical modeling and software design in order to develop various numerical techniques in the context of high-performance computing.

He started his career in industry as a Consultant in a spatial project for the CNES, and joined the Delair team in late 2018 as a Computer Vision and Machine Learning Engineer. As a member of the Delair Data Team, he worked in several projects involving image and point cloud manipulation using multiple techniques involving deep learning applied to computer vision. He is currently the Head with the Computer Vision and Data Science Department. He has authored and co-authored in various international reviews and presented in multiple conferences.

Guy Le Besnerais received the graduate degree in physics from ENSTA, Palaiseau, France, in 1989, the Ph.D. degree from Université Paris-Sud, Paris, France, in 1993, and the grade of "Habilitation à diriger les recherches" (HDR) from the Université Paris-Saclay, Saclay, France, in 2008.

Since 1994, he has been with the Information Processing and Systems Department, ONERA, Palaiseau, France, The French Aerospace Lab, where he became a Senior Scientist in 2012. His research interests include methodological and algorithmic studies for inverse and low-level vision problems resolution, for geometrical vision such as 3-D and (ego)motion estimation from video and for object recognition and image understanding. His works found several applications in video processing, satellite and airborne remote sensing, imagin-based metrology, and codesign of hybrid sensor+processing systems.