Hierarchical Instance Mixing across Domains in Aerial Segmentation

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 December 2022, accepted 6 February 2023, date of publication 8 February 2023, date of current version 13 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243475

RESEARCH ARTICLE

Hierarchical Instance Mixing Across Domains in Aerial Segmentation

EDOARDO ARNAUDO[1,2], ANTONIO TAVERA[1], (Graduate Student Member, IEEE), CARLO MASONE[1], (Member, IEEE), FABRIZIO DOMINICI[2], AND BARBARA CAPUTO[1]

[1]Dipartimento di Automatica e Informatica (DAUIN), Politecnico di Torino, 10129 Turin, Italy
[2]AI, Data and Space (ADS), Links Foundation, 10138 Turin, Italy

Corresponding author: Edoardo Arnaudo (edoardo.arnaudo@polito.it)

This work was supported by the Horizon 2020 Project Structured Approaches for Forest fire Emergencies in Resilient Societies (SAFERS) under Grant 869353.

ABSTRACT We investigate the task of unsupervised domain adaptation in aerial semantic segmentation observing that there are some shortcomings in the class mixing strategies used by the recent state-of-the-art methods that tackle this task: 1) they do not account for the large disparity in the extension of the semantic categories that is common in the aerial setting, which causes a domain imbalance in the mixed image; 2) they do not consider that aerial scenes have a weaker structural consistency in comparison to the driving scenes for which the mixing technique was originally proposed, which causes the mixed images to have elements placed out of their natural context; 3) source model used to generate the pseudo-labels may be susceptible to perturbations across domains, which causes inconsistent predictions on the target images and can jeopardize the mixing strategy. We address these shortcomings with a novel aerial semantic segmentation framework for UDA, named HIUDA, which is composed of two main technical novelties: firstly, a new mixing strategy for aerial segmentation across domains called Hierarchical Instance Mixing (HIMix), which extracts a set of connected components from each semantic mask and mixes them according to a semantic hierarchy and secondly, a twin-head architecture in which two separate segmentation heads are fed with variations of the same images in a contrastive fashion to produce finer segmentation maps. We conduct extensive experiments on the LoveDA benchmark, where our solution outperforms the current state-of-the-art.

INDEX TERMS Computer vision, deep learning, image processing, remote sensing, semantic segmentation, unsupervised domain adaptation.

I. INTRODUCTION

Semantic segmentation is a well-known computer vision task that aims to predict the semantic category of each individual pixel in an image from a predefined set of possible labels. Such fine-grained scene understanding has numerous applications in aerial robotics and remote sensing, where it is used to perform road extraction and building detection [1], [2], [3], to classify land cover [3], [4], [5], [6], to estimate damages caused by wildfires [7], to assess deforestation [8] and to classify agricultural patterns [9].

In the field of aerial semantic segmentation there have been numerous advances based on the use of deep learning models [2], [3], [6], [10], [11] trained on large public datasets with annotated images [9], [12], which have led to remarkable levels of performance. However, this performance generally does not carry over when these models are set to operate on images that come from a distribution (*target domain*) different from the data experienced during training (*source domain*). In principle this loss in performance could be recovered by fine-tuning the source model on large quantities of labeled images collected from the target domain, but in practice this is generally not an option because generating pixel-wise annotations for semantic segmentation is extremely costly [13]. Thus, in this paper we tackle the problem of unsupervised domain adaptation (UDA) for aerial semantic segmentation assuming to have available at training time both a set of labeled images from the source domain and a collection of unlabeled images from the target domain. This problem is widely studied in the literature, and a prominent

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan-Li Sun.

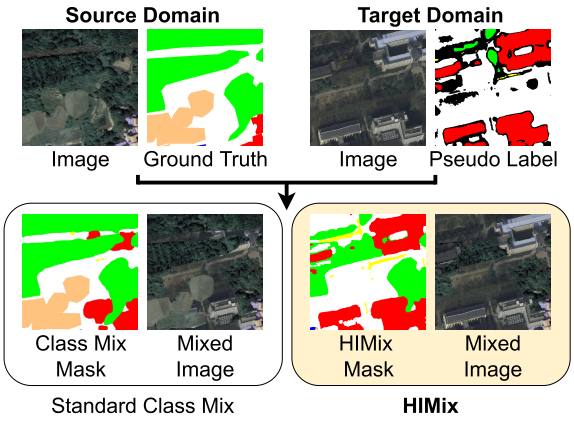

FIGURE 1. Class Mix overlays classes of the source domain onto the target without taking into account the semantic hierarchy of the visual elements. As a result, it generates erroneous images that are detrimental to Unsupervised Domain Adaptation training in the aerial scenario. Instead, our HIMix extracts instances from each semantic label and then composes the mixing mask after sorting the extracted instances based on their pixel count. This mitigates some artifacts (e.g. partial buildings) and improves the balance of the two domains.

solution to address it is to combine domain mixing solutions with self-supervision [14], [15], [16], [17]. Namely, the source model is used to generate semantic predictions (*pseudo-labels*) on the unlabeled target data, and then the labeled source images and the pseudo-labeled target images are mixed to create artificial images with elements from both domains. Training on these mixed samples encourages the model to learn domain-agnostic features. In particular, both DACS [15] and DAFormer [16], two recent state-of-the-art methods in aerial semantic segmentation, rely on ClassMix [18], a mixing strategy originally created for driving scenes that generates the composite image by randomly selecting half of the semantic classes from the source image and pasting them onto the target one (see Fig. 1 left). However, we argue that such a self-supervised mixing solution has a few shortcomings that make it inadequate for aerial semantic segmentations:

- **Domain imbalance:** Segmentation-oriented aerial datasets are often characterized by categories with vastly different extensions, some exposing only a few pixels (e.g., *cars*) and others occupying large portions of the whole image (e.g., *forest*). This disparity in raw pixel counts between classes may be detrimental for an effective domain adaptation through class mixing, because the composition may be skewed towards on of the two domains, depending on how the classes have been sampled (see Fig. 1 left), which in turn leads to a poor alignment of the features.
- **Out-of-context instances:** The mixing strategies used in aerial segmentation, such as ClassMix [18], were originally developed for applications in driving scenes. However, the scenes captured by a front-facing camera onboard a car have a consistent structure, with the street at the bottom, the sky at the top, sidewalks and buildings at the sides, etc. This structure is preserved also across domains, as in the classic Synthia [19] → CityScapes [13] setting. Thus, when copying objects from a driving scene onto another one they are likely to end up in a reasonable context. This is not true for aerial images, where there is no consistent semantic structure (see Fig. 1 left).
- **Pseudo-labels:** The effectiveness of the semi-supervised mixing strongly depends on the accuracy of the pseudo-labels generated on the target images during training. However, the source model used to generate the pseudo-labels may be susceptible to perturbations across domains, thus leading to inconsistent predictions on the target images and possibly jeopardizing the domain mixing strategy.

In this paper, we propose a new framework for unsupervised domain adaptation in aerial semantic segmentation, called **Hierarchical Instance Mixing for Unsupervised Domain Adaptation** (**HIUDA**). HIUDA addresses the aforementioned shortcomings in current domain mixing strategies by introducing two technical novelties:

- A new mixing strategy for aerial segmentation across domains, called **Hierarchical Instance Mixing** (HIMix). HIMix extracts from each semantic mask a set of connected components, akin to instance labels. The intuition is that aerial tiles often present very large stretches of land, divided into instances (e.g., forested areas separated by a road). HIMix randomly selects from the individual instances a set of layers that will compose the binary mixing mask. This helps to mitigate the pixel imbalance between source and target domains in the artificial image. Afterward, HIMix composes these sampled layers by sorting them based on the observation that there is a semantic hierarchy in the aerial scenes (e.g., cars lie on the road and roads lie on stretches of land). We use the pixel count of the instances to determine their order in this hierarchy, placing smaller layers on top of larger ones. This hierarchical composition helps to mitigate the occurrence of semantic instances pasted in an unreasonable context (e.g. cars in the water) and it also reduces the bias towards those categories with larger surfaces in terms of pixels as they are placed below the other layers of the mask (see Fig. 1 right).
- A new **twin-head UDA architecture** in which two separate segmentation heads are fed with contrastive variations of the same images to improve pseudo-label confidence and make the model more robust and less susceptible to perturbations across domains, inevitably driving the model towards augmentation-consistent representations.

We evaluate HIUDA on the LoveDA benchmark [12], the only dataset designed for evaluating unsupervised domain adaptation in aerial segmentation, where we exceed the current state-of-the-art. We further provide a comprehensive

ablation study to assess the impact of the proposed solutions. The code will be made available to the public to foster the research in this field.

The rest of this paper is organized as follows. In Sec. II we provide an analysis on the literature in this field and discuss the related works. Then, in Sec. III we formally state the problem setting and we describe in detail the proposed HIUDA method, focusing first on the novel HIMix mixing strategy and then on the twin-head architecture. In Sec. IV we proceed to present the experimental validation of the proposed method, describing first the datasets used and the implementation details, and then discussing the results. Lastly, in Sec. V we provide some conclusive remarks regarding the limitations of the method as well as possible directions for future research.

II. RELATED WORK

A. AERIAL SEMANTIC SEGMENTATION

Current semantic segmentation methods mostly rely on convolutional encoder-decoder architectures [20], [21], [22], [23], but the recent breakthroughs of vision Transformers introduced new effective encoder architectures such as ViT [24], Swin [25] or Twins [26], as well as end-to-end segmentation approaches such as Segmenter [27] and SegFormer [28]. Concerning the application to aerial images, despite the comparable processing pipeline as in other settings, there are peculiar challenges that demand for specific solutions. Firstly, aerial and satellite data often include multiple spectra besides the visible bands, which can be leveraged in different ways, such as including them as extra channels [9] or adopting multi-modal encoders [2]. Yet, this is not the case for the LoveDA dataset used in this manuscript. Visual features represent another major difference: unlike other settings, aerial scenes often display a large number of entities on complex backgrounds, with wider spatial relationships. In this case, attention layers [29] or relation networks [30] are often employed to better model long-distance similarities among pixels. Transformers are in this case a natural choice, considering their attention-based architecture which is capable of extracting long-range relations. Another distinctive trait of aerial imagery is the top-down point of view and the lack of reference points that can be observed in natural images (e.g., sky is always on top). This can be exploited to produce rotation-invariant features using ad-hoc networks [11], [31], or through regularization [32]. Considering the UDA setting, feature consistency across domains is crucial for an effective training. Forcing the output to be invariant not only to rotation but also to other geometric and photometric transformations, can be extremely beneficial for the generalization abilities of the final model. Lastly, aerial images are characterized by disparities in class distributions, since these include small objects (e.g. cars) and large stretches of land. This pixel imbalance can be mitigated with sampling and class weighting [16], or ad-hoc loss functions [33]. In the context of aerial images, these solutions may not be enough to tackle this issue, given the wide range of scales. In this work, we attempt to further balance this disparity by ensuring that smaller objects always appear on top, through the hierarchical mixing approach. While approximating a top-down hierarchy, this method ensures that large surfaces do not overshadow less represented categories, allowing for fairer training.

B. DOMAIN ADAPTATION

Domain Adaptation (DA) is the task of attempting to train a model on one domain while adapting to another. The main objective of domain adaptation is to close the *domain shift* between these two dissimilar distributions, which are commonly referred to as the source and target domains. The initial DA techniques proposed in the literature attempt to minimize a measure of divergence across domains by utilizing a distance measure such as the Maximum Mean Discrepancy (MMD) [34], [35], [36]. Another popular approach to DA in Semantic Segmentation is adversarial training [37], [38], [39], [40], which involves playing a min-max game between the segmentation network and a discriminator. This latter is responsible for discriminating between domains, whereas the segmentation network attempts to trick it by making features of the two distributions identical. However, when attempting to align the features of two domains, it is common for samples with different semantic labels to be mixed together. This can often lead to a class mismatch between the two domains. Other approaches, such as [41], [42], [43], employ image-to-image translation algorithms to generate target pictures styled as source images or vice versa, thus not tackling the issue of differences in texture and semantic content of classes between the two domains. More recent methods like [44], [45], [46] use self-learning techniques to generate fine pseudo-labels on target data to fine-tune the model. Nonetheless, these methods directly select pseudo-labels with high prediction confidence, which can lead to the model being biased towards easy classes and negatively impacting its ability to transform the hard classes.

Novel methods like [15], [16] combine self-training with the class mix to reduce low-quality pseudo-labels caused by domain shifts among the different distributions. These mixing algorithms are very effective on data with a consistent semantic organization of the scene, such as in self-driving scenes [13], [47]. In these scenarios, naively copying half of the source image onto the target image increases the likelihood that the semantic elements will end up in a reasonable context. This is not the case with aerial imagery (see Fig. 1). HIMix not only mitigates this problem, but it also reduces the bias towards categories with larger surfaces.

III. METHODOLOGY

A. PROBLEM STATEMENT

In this paper, we focus on the task of aerial semantic segmentation in the context of unsupervised domain adaptation (UDA). UDA is a form of transfer learning that involves adapting a model to a new scenario, or target domain, using only labeled data from a source domain and unlabeled data

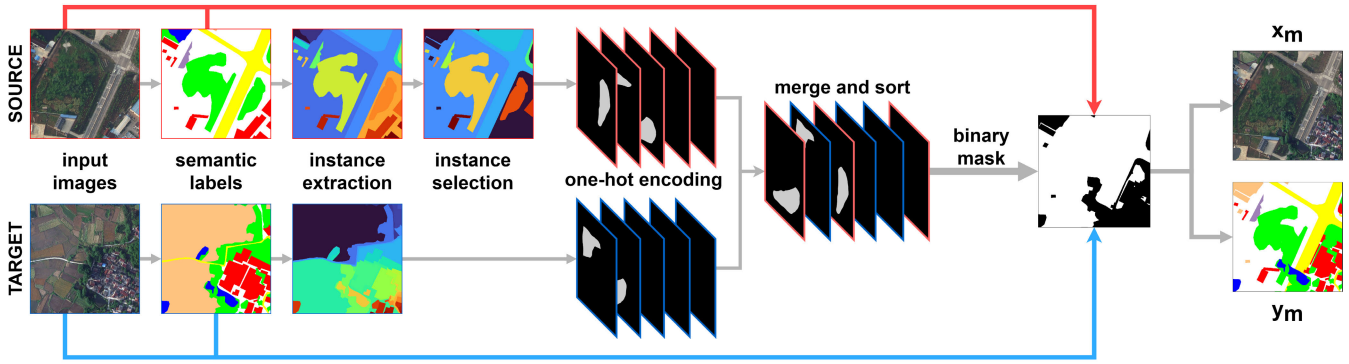

FIGURE 2. HIMix operates by (i) extracting the connected components from the source label and target pseudo-label, (ii) selecting uniformly which instances should be mixed from S, (iii) merging source and target instances hierarchically based on instance size (smaller ones on top), and (iv) producing a binary mask M to construct the final blended image x_m and its label y_m.

from the target domain. By studying aerial semantic segmentation in the context of UDA, we aim to improve the performance of this task in new scenarios where labeled data may be limited. More specifically, let us define as $\mathcal{X}$ the set of RGB images constituted by the set of pixels $\mathcal{I}$, and as $\mathcal{Y}$ the set of semantic masks associating a class from the set of semantic classes $\mathcal{C}$ to each pixel $i \in \mathcal{I}$. We have two sets of data accessible at training time: (i) a set of annotated images from the source domain, denoted as $X_s = \{(x_s, y_s)\}$ with $x_s \in \mathcal{X}$ and $y_s \in \mathcal{Y}$; (ii) a set of N_t unlabeled images from the *target* domain, denoted as $X_t = \{(x_t)\}$ with $x_t \in \mathcal{X}$.

The goal is to find a parametric function f_θ that maps an RGB image to a pixel-wise probability, i.e., $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^{|\mathcal{I}|\times|\mathcal{C}|}$, and evaluate it on unseen images from the target domain. In the following, we indicate the model output in a pixel i for the class c as p_i^c, i.e., $p_i^c(x) = f_\theta(x)[i, c]$. Following standard practices in literature [16], [20], [21], [48], the parameters θ are tuned to minimize a categorical cross-entropy loss defined as:

$$L_{\text{seg}}(x, y) = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{c\in\mathcal{C}} y_i^c \log(p_i^c(x)), \tag{1}$$

where y_i^c represents the ground truth annotation for the pixel i and class c. While alternative functions or a combination of them, such as cross-entropy and Dice loss [12], could be adopted in the aerial domain, this work concentrates on the UDA task. In this context, the objective function is not the main focus, and the cross-entropy provides a fair comparison with other approaches [16].

B. FRAMEWORK

The remainder of this section describes a new end-to-end trainable UDA framework, named HIUDA, based on the use of target pseudo-labels. To better align domains, we construct artificial images using our HIMix strategy (III-C), which generates mixed images exploiting the instances produced both from the source ground truth and the target pseudo-label. Rather than using a secondary teacher network derived from the student as an exponential moving average as in [15], [16], we propose a twin-head architecture (III-D) with two separate decoders trained in a contrastive fashion to provide finer target pseudo labels.

C. HIERARCHICAL INSTANCE MIXING

Given the pairs (x_s, y_s) and $(x_t, \hat{y}_t)$, where $\hat{y}_t = f_\theta(x_t)$ are the pseudo-labels computed from the model prediction on the target domain, the purpose of the mixing strategy is to obtain a third pair, namely (x_m, y_m), whose content is derived from both source and target domains using a binary mask M. While techniques based on ClassMix have been successfully applied in many UDA settings, we discover that the same may not be optimal in the aerial scenario since it superimposes parts of the source domain onto the target without taking into consideration their semantic hierarchy (e.g., cars appear on top of roads, not vice versa). In contrast, we propose a Hierarchical Instance Mixing strategy (HIMix), which is composed of two subsequent steps: (i) *instance extraction* and (ii) *hierarchical mixing*.

1) INSTANCE EXTRACTION

Aerial tiles often present uniform land cover features, with many instances of the same categories in the single image. In the absence of actual instance labels, this peculiarity can be exploited to separate semantic annotations into connected components. Here a connected component is a set of pixels that have the same semantic label and such that for any two pixels in this set there is a path between them that is entirely contained in the same set. Fig. 2 illustrates an example of this process, with a forest that is separated in two instances by a road. This increases the number of regions which can be randomly selected for the mixing phase, thus mitigating the pixel unbalance in the final mixed sample between source and target domains. Note that this procedure is applied to the concatenation of source and target label.

2) HIERARCHICAL MIXING

We observe that instances in aerial imagery have an inherent hierarchy that is dictated by their semantic categories. In other words, land cover categories such as *barren* or *agricultural* frequently appear in the background w.r.t. smaller instances

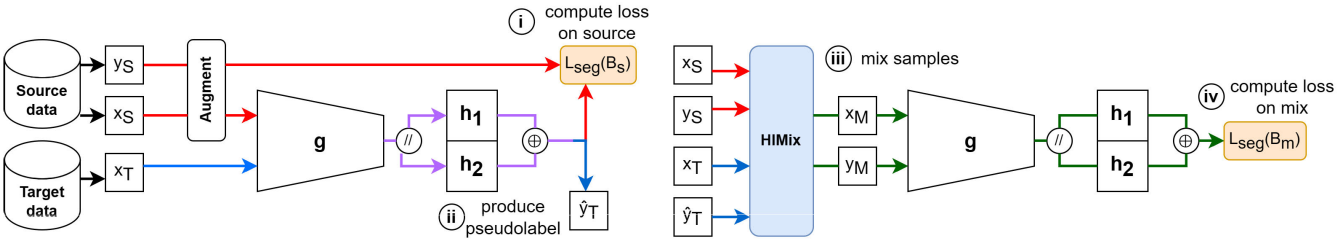

FIGURE 3. Our framework training: (i) standard training is carried out on the source domain, on samples composed of *source image* (x_S) and *target image* (y_S, through the segmentation loss $L_{seg}(B_S)$. (ii) Pseudolabels $\hat{y}_T$ are generated from target the target image (x_T), encoded by the shared backbone g, through majority voting between each segmentation head output(h_1, h_2). (iii) Source and target samples are mixed together through *HIMix*, producing a new pair of samples x_M and y_M. (iv) Last, the segmentation loss $L_{seg}(B_M)$ is computed on mixed pairs.

such as *roads* or *buildings*. The mixing step follows this hierarchy when combining the instances from source and target, and it is illustrated in Fig. 2. First, both sets of instance labels are encoded into a one-hot representation,[1] so that each component yields its own mask layer. Then both stacks of layers are merged together and sorted by their pixel count, with the larger layers on the bottom. Finally, a reduction from top to bottom projects the 3D tensor into a 2D binary mask M, where positive values indicate *source* pixels, and null values indicate *target* pixels.

D. TWIN-HEAD ARCHITECTURE

State-of-the-art, self-training UDA strategies, such as [16], use *teacher-student* networks to improve the consistency of the pseudo-labels. Albeit dealing with consistency in time, teacher-based approaches do not directly cope with geometric or stylistic consistency. We propose a twin-head segmentation framework to directly address this, providing more consistent pseudo-labels and outperforming the standard tested methodologies, as shown in the ablation study IV-C. Our architecture (see Fig. 3) comprises a shared encoder g, followed by two parallel and lightweight segmentation decoders, h_1 and h_2. Training is carried out end to end, exploiting annotated source data and computing pseudo-labels from target images online, as detailed hereinafter.

1) SOURCE TRAINING

With the purpose of driving the model towards augmentation-consistent representations, we feed the two heads with variations of the same source image in a contrastive fashion. More specifically, at each iteration we consider a random sequence of geometric augmentations T_g (*horizontal flipping*, *rotation*) and photometric augmentations T_p (*color jitter*). At each iteration, given these sampled transformations, a source image x_s and its ground-truth label, we compute their augmented counterparts as $\tilde{x}_s = T_p(T_g(x_s))$, and $\tilde{y}_s = T_g(y_s)$. After this operation, the full augmented pair $B_s = (\text{concat}(x_s, \tilde{x}_s), \text{concat}(\ y_s, \tilde{y}_s))$, where concat$(\cdot)$ is the concatenation function, is first forwarded to the shared encoder module g, producing a set of features. The latter, containing information derived from the images and its augmented variants, are split and forwarded to the two parallel heads, effectively obtaining two comparable outputs, $h_1(g(x_s))$ and $h_2(g(\tilde{x}_s))$. A standard cross-entropy loss, as shown in Eq. 1, is computed on both segmentation outputs. Working independently on different variations of the same images, the two heads can evolve in different ways while trying to minimize the same objective function. Using the same encoder yields a more robust, contrastive-like feature extraction that is less susceptible to perturbations. This is essential for producing more stable and precise pseudo-labels.

2) MIX TRAINING

The twin-head architecture is expressly designed to generate more refined pseudo-labels. Given an unlabeled target image x_t, we compare the probabilities $\sigma(h_1(g(x_t)))$ and $\sigma\ (h_2(g(x_t)))$ obtained by forwarding the image to both heads and passing them through a Softmax function, and select the maximum value between the two. Once p_i^c is derived, the pseudo-label $\hat{y}_t$ necessary for the mixing strategy is generated for each head output through:

$$\hat{y}_t^{(i,c)} = [c = \underset{c}{\text{argmax}}\ p_i^c(x_t)] \tag{2}$$

At this point, the mixed pairs of inputs can be computed through HIMix, as described in previous sections and detailed in Algorithm 1, obtaining (x_m, y_m) as a composition of the source and target samples. Similar to source training, an augmented pair $B_m = (\text{concat}(x_m, \tilde{x}_m),\ \text{concat}(\ y_m, \tilde{y}_m))$ is computed through geometric and photometric transformations, then fed to the model to compute $L_{seg}(B_m)$. To reduce the impact of low-confidence areas, a pixel-wise weight map w_m is generated. Similar to [15], [16], w_m is computed as the percentage of valid points above threshold. Formally, for each pixel i:

$$w_m^i = \begin{cases} 1, & i \in y_s \\ \dfrac{m_\tau}{|\mathcal{I}|}, & i \in \hat{y}_t \end{cases} \tag{3}$$

where m_τ represents the Max Probability Threshold [49] computed over pixels belonging to the pseudo-label as

[1]The one hot representation is an index map where each pixel indicates the index of the class it belongs to.

follows:

$$m_\tau^i = \mathbb{1}_{[argmax_c p_i^c(x_t) > \tau]}, \tag{4}$$

In practice, each pixel of the mixed label is either weighted as 1 for regions derived from the source domain, or by a factor obtained as the number of pixels above the confidence threshold, normalized by the total amount of pixels. Note that, during all of these computations, the gradients are not propagated. The overall HIUDA training procedure is detailed in Algorithm 1.

Algorithm 1 HIUDA Training Procedure

Initialize:
Model $f_\theta : \mathcal{X} \rightarrow \mathbf{R}^{|\mathcal{I}| \times |\mathcal{Y}|}$ with encoder g and twin heads h_1, h_2;
Input: $\mathcal{X}_s$ source domain with N_s pairs (x_s, y_s), $x_s \in \mathcal{X}$, $y_s \in \mathcal{Y}$ and semantic classes $\mathcal{C}$;
$\mathcal{X}_t$ target domain with N_t images x_t, lacking ground truth labels;
Output: $y = \{\text{argmax}_{c \in \mathcal{Y}} p_i^c\}_{i=1}^N$, where p_i^c the model prediction of pixel i for class c and $\mathcal{Y}$ the label space;
while *epoch in max_epochs* **do**
  while x_s, y_s, x_t *in* $\mathcal{X}_s \times \mathcal{X}_t$ **do**
    Train on *source* $\mathcal{X}_s$
      // Compute augmented source batch
      $B_s = (\text{concat}(x_s, \tilde{x}_s), \text{concat}(y_s, \tilde{y}_s))$;
      // Train f_θ on source labels with $L_{seg}(B_s)$
    end
    Mix *source* and *target* pairs
      // Compute pseudo-labels via majority
      // voting $\hat{y}_t = max\ (h_1(g(x_t)), (h_2(g(x_t))))$;
      // Extract source instance labels
      $i_s = CCL(y_s)$ with instances $\in K_s$;
      // Extract target instance pseudo-labels
      $i_t = CCL(\hat{y}_t)$ with instances $\in K_t$;
      // Compute one-hot encoded labels,
      // sorted by pixel size as:
      $1_m = sorted\ \left(\text{concat}(1_{K_s}(i_s), 1_{K_t}(i_t))\right)$;
      // Reduce z axis to 2D indexed mask
      $m = argmax_z 1_m(i, j, z)$;
      // Binarize mask
      $\forall i, j \in m,\ M = \begin{cases} 1 & if\ m(i,j) \in K_s \\ 0 & if\ m(i,j) \in K_t \end{cases}$;
      // Compute mixed image and labels as:
      $x_m = M \odot x_s + (1 - M) \odot x_t$;
      $y_m = M \odot y_s + (1 - M) \odot \hat{y}_t$;
      // Compute w_m as in Eq. 3
    end
    Train on *mixed* $\mathcal{X}_m$ **pairs**
      // Compute augmented mixed batch
      $B_m = (\text{concat}(x_m, \tilde{x}_m), \text{concat}(y_m, \tilde{y}_m))$;
      // Train f_θ on mixed samples with
      // $L_{seg}(B_m)$, weighted by w_m
    end
  end
end

IV. EXPERIMENTS

A. TRAINING DETAILS

We assess the performance of our approach on the LoveDA dataset [12]. According to that benchmark, we conduct two series of unsupervised domain adaptation experiments: *rural→urban* and *urban→rural*. We have a separate set of data for training and testing on the target domain. The target training set is unlabeled, while the testing set has ground truth labels that allow us to measure the performance of the model.

1) DATASET

To our knowledge, the LoveDA dataset [12] is the only open and free collection of land cover semantic segmentation images in remote sensing explicitly designed for UDA. Both urban and rural areas are included in the training, validation, and test sets. Data is gathered from 18 different administrative districts in China. The urban training set has 1156 images, while the rural training set contains 1366 images. Each image is supplied in a tiled format of 1024 × 1024 pixels annotated with seven categories.

2) METRIC

Following [12] we use the averaged Intersection over Union (mIoU) metric to measure the accuracy of all the experiments conducted.

3) BASELINES

We compare HIUDA to various cutting-edge UDA methods. The first baseline we consider is the Source Only model, which is a network that has only been trained using the source dataset. We look at MMD's [36] original metric-based methodology. Then, we compare two alternative UDA approaches: the adversarial training strategy, with AdaptSegNet [37], FADA [39], CLAN [38], and TransNorm [50], and the self-training technique, with CBST [45], PyCDA [44], IAST [46], DACS [15], and DAFormer [16].

4) IMPLEMENTATION

To implement HIUDA we leverage the *mmsegmentation* framework, that is based on PyTorch. We train each experiment on an NVIDIA Titan GPU with 24 GB of RAM. We refer to DAFormer [16] for the architecture and configuration of hyperparameters. We use the MiT-B5 model [28] pretrained on ImageNet as the encoder of our method, while the segmentation decoder module corresponds to the SegFormer head [28]. We train on every setting for 40k iterations using AdamW as the optimizer. The learning rate is set to 6×10^{-5}, weight decay of 0.01, betas to (0.9, 0.99). We also adopt a polynomial decay with a factor of 1.0 and warm-up for 1500 iterations. To cope with possible variations, every experiment presented has been obtained as the average over three seeds {0, 1, 2}. Training is performed on random crops,

TABLE 1. Urban→Rural experiments. Experiments marked with * were replicated using the original method.

Method	Backg.	Building	Road	Water	Barren	Forest	Agric.	mIoU
Source Only	24.2	37.0	32.6	49.4	14.0	29.3	35.7	31.7
MCD [36]	25.6	44.3	31.3	44.8	13.7	33.8	26.0	31.4
AdaptSeg [37]	26.9	40.5	30.7	50.1	17.1	32.5	28.3	32.3
FADA [39]	24.4	33.0	25.6	47.6	15.3	34.4	20.3	28.7
CLAN [38]	22.9	44.8	26.0	46.8	10.5	37.2	24.5	30.4
TransNorm [50]	19.4	36.3	22.0	36.7	14.0	40.6	03.3	24.6
PyCDA [44]	12.4	38.1	20.5	57.2	18.3	36.7	41.9	32.1
CBST [45]	25.1	44.0	23.8	50.5	08.3	39.7	49.7	34.4
IAST [46]	30.0	49.5	28.3	64.5	02.1	33.4	61.4	38.4
DACS* [15]	20.1	50.5	35.9	60.6	09.9	35.4	17.5	32.9
DAFormer* [16]	29.5	57.9	41.8	67.1	07.6	35.3	48.1	41.0
HIUDA	31.5	59.6	51.5	68.1	08.2	37.4	53.9	**44.3**

TABLE 2. Rural→Urban experiments. Experiments marked with * were replicated using the original method.

Method	Backg.	Building	Road	Water	Barren	Forest	Agric.	mIoU
Source Only	43.3	25.6	12.7	76.2	12.5	23.3	25.1	31.3
MCD [36]	43.6	15.4	12.0	79.1	14.3	33.1	23.5	31.5
AdaptSeg [37]	42.4	23.7	15.6	82.0	13.6	28.7	22.1	32.6
FADA [39]	43.9	12.6	12.8	80.4	12.7	32.8	24.8	31.4
CLAN [38]	43.4	25.4	13.8	79.3	13.7	30.4	25.8	33.1
TransNorm [50]	33.4	05.0	03.8	80.8	14.2	34.0	17.9	27.7
PyCDA [44]	38.0	35.9	45.5	74.9	07.7	40.4	11.4	36.3
CBST [45]	48.4	46.1	35.8	80.1	19.2	29.7	30.1	41.3
IAST [46]	48.6	31.5	28.7	86.0	20.3	31.8	36.5	40.5
DACS* [15]	46.0	31.6	33.8	76.4	16.4	29.3	27.7	37.3
DaFormer* [16]	49.2	47.7	55.2	86.6	16.5	39.5	30.8	46.5
HIUDA	49.3	55.0	55.4	86.0	17.1	41.2	36.9	**48.7**

TABLE 3. Ablation study on our HIUDA framework which comprises the twin-head architecture and HIMix strategy.

ID	Twin Head	Class Mix	Instance Mix	Hierarc. Mix	mIoU U2R	mIoU R2U
1		✓			41.0 ± 0.33	46.5 ± 0.41
2			✓	✓	43.4 ± 0.76	47.6 ± 0.10
3	✓	✓			42.9 ± 0.35	47.1 ± 0.34
4	✓		✓		43.2 ± 0.35	47.4 ± 0.16
5	✓		✓	✓	**44.3** ± 0.39	**48.7** ± 0.06

by augmenting data through random resizing in the range [0.5, 2.0], horizontal and vertical flipping, and rotation of 90 degrees with probability $p = 0.5$, together with random photometric distortions (i.e., brightness, saturation, contrast, and hue). As [15], [16], we set $\tau = 0.968$ in 4. The final inference on the test set is instead performed on raw images without further transformations. HIUDA trains in approximately 11 hours on an NVIDIA Titan GPU.

B. RESULTS

1) Urban→Rural

The results for this set of experiments are reported in Tab. 1. They corroborate the complexity of the task due to a strong and inconsistent class distribution in the source domain, which is dominated by urban scenes with a mix of buildings and highways but few natural items. This causes a negative transfer to the target domain since both adversarial strategies and self-training procedures achieve overall performance equivalent to, if not worse than, the Source Only model. Specifically, when we evaluate the best performing Adversarial Training technique, which is represented by CLAN, we gain just a +1.8 improvement over the Source Only model. Self-training approaches have shown to be the most effective. DACS, which introduces the class mix strategy, improves the *Source Only* model by +1.2, while DAFormer, which uses a Transformer backbone and the same class mix strategy as DACS, outperforms the *Source Only* model by +9.3. HIUDA, which combines both the twin-head architecture and the innovative class mix, outperforms the *Source Only* model by a wide margin of +12.6 and it exceeds its closest competitor (DAFormer) by +3.3. HIUDA exhibits its ability to boost rural and underrepresented classes, such as *agriculture*, as also evidenced by qualitative results in Fig. 4. In comparison to DACS and DAFormer, our technique recognizes and classifies better contours and classes, such as *water*, despite their under-representation in the source domain. This is also true in common categories with different visual features, such as *road*, which can appear in paved and unpaved variants.

2) Rural→Urban

The results for this set of experiments are summarized on Tab. 2. The source domain in this scenario is dominated by large-scale natural objects and a few man-made samples. Nonetheless, the models under consideration are capable of effectively transferring knowledge even in these underrepresented categories. Self-learning approaches outperform adversarial methods, getting an average boost of +9.1 over the Source Only model, whereas adversarial training methods achieve comparable accuracy. In terms of mIoU, the two best performing self-training models and our closest competitor surpass the Source Only model by +6.0 and +15.2, respectively. In comparison, HIUDA gains a +17.4 boost over the Source Only model, outperforming DACS and DAFormer by +11.4 and +2.2, respectively. In this case, the qualitative results in Fig. 4 support the superior ability of HIUDA to discern between rural and urban classes. While DACS does not recognize *buildings* and DAFormer misclassifies parts of them as *agricultural* terrain, our model demonstrates its effectiveness in minimizing the bias towards those categories with larger surfaces, providing results close to the ground truth.

C. ABLATION

1) TWIN-HEAD AND HIMix

To demonstrate the effectiveness of the twin-head architecture, we compare it to the traditional single-head structure, which generates pseudo-labels using a secondary teacher network derived from the student as an exponential moving average. This study also demonstrates the potential of the HIMix when paired with traditional single-head training. For both settings, we conduct an extensive ablation study considering the MiT-B5 [28] as the backbone, and we report the results in Tab. 3.

The twin-head design paired with the Standard Class Mix (line 3) performs better than the single-head architecture (line

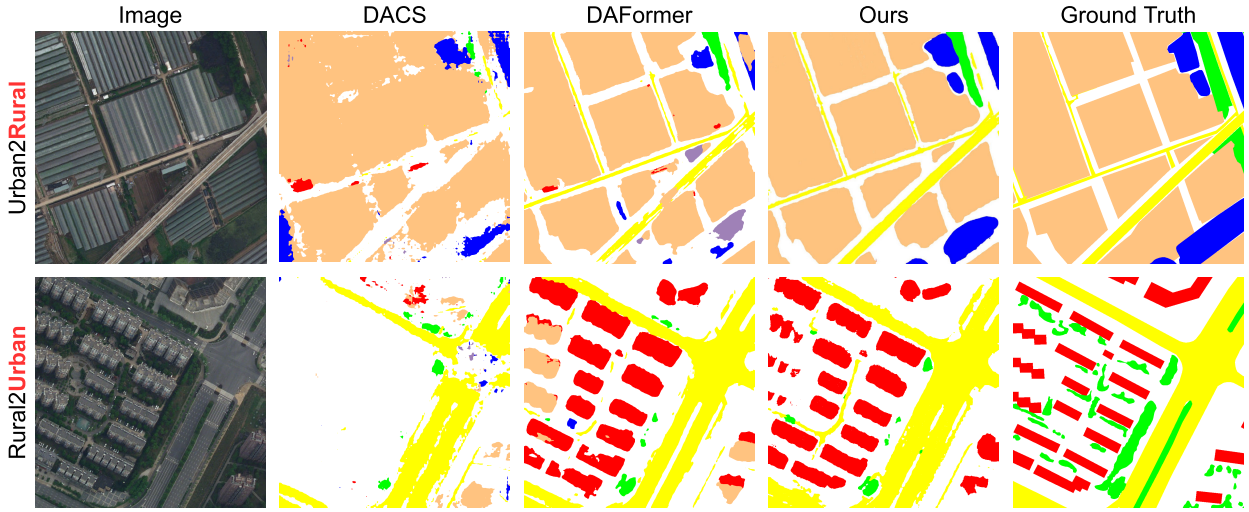

FIGURE 4. Qualitative results in the two settings, ***Urban→Rural*** and ***Rural→Urban***, after testing on target domain.

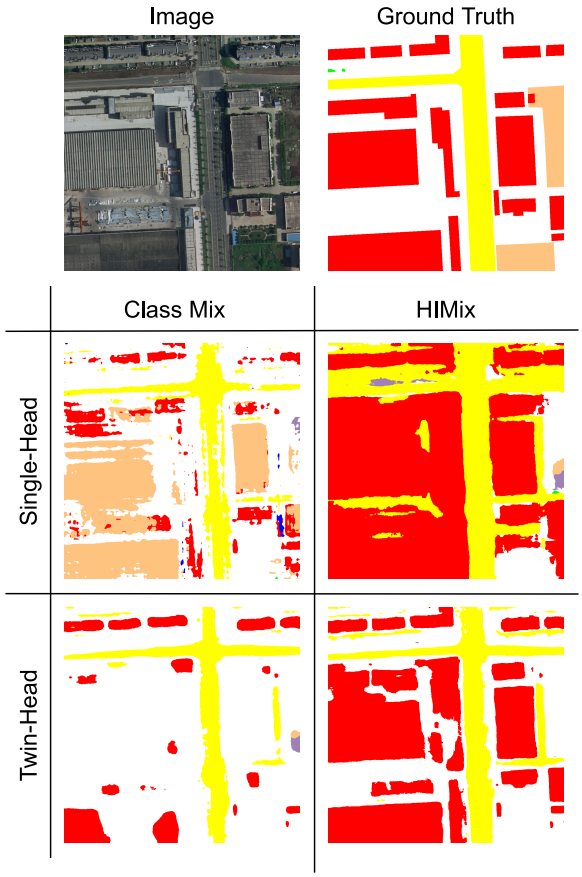

FIGURE 5. Qualitative comparison of Single or Twin-Head architectures using Standard Class Mix or our HIMix.

1), implying that our solution is better at providing finer pseudo-labels with correct class segmentation, as shown also in the first column of Fig. 5.

HIMix increases recognition performance even when paired with a single-head architecture (line 2), particularly for categories with a lower surface area in terms of pixels, which are placed below those with larger surfaces when using the Standard Class Mix. That is why, in the top-left image of Fig. 5, the model is unable to grasp their semantics effectively and erroneously classifies *building* as *agricultural* pattern. In comparison, HIMix can accurately distinguish *buildings* (top-right picture in Fig. 5) even though the prediction has poorly defined contours.

The best results are obtained when the twin-head ability to provide an enhanced segmentation map is combined with the HIMix ability to maintain a correct semantic structure (line 5). This combination yields the best results in terms of accuracy and level of detail in the segmentation map, as shown in the bottom-right image of Fig. 5.

We finally ablate the different components of our HIMix to assess each term's contribution to overall performance (lines 4-5). The Hierarchical Mixing always increases the Instance Extraction by +1.1 and +1.3 in the two Urban→Rural and Rural→Urban scenarios, respectively.

V. CONCLUSION

We investigated the problem of Unsupervised Domain Adaptation in aerial Semantic Segmentation, showing that the peculiarities of aerial imagery, principally the lack of structural consistency and a significant disparity in semantic class extension, must be taken into consideration. We addressed these issues by developing HIUDA, an end-to-end trainable UDA framework comprising two main contributions. First, a novel domain mixing method that consists of two parts: an instance extraction that chooses the connected components from each semantic map and a hierarchical mixing that sorts and fuses the instances based on their pixel counts. Second, a twin-head architecture that produces finer pseudo labels for the target domain, improving the efficacy of the domain mixing. We demonstrated the effectiveness of HIUDA with a comprehensive set of experiments in the LoveDA benchmark, outperforming the previous state-of-the-art by 3.3 in the urban→rural and 2.2 in the rural→urban settings.

1) LIMITATIONS

Despite the excellent results, we observed that our solution has worse performance than the Source Only model in the *barren* class, particularly in the Urban→Rural scenario. This is possibly due to the large disparity in absolute pixels count between source and target domains in this category.

2) FUTURE WORKS

In the future, we plan to evaluate lighter segmentation heads and other contrastive techniques to accelerate overall training and improve performance, particularly on specific semantic classes.

ACKNOWLEDGMENT

(Edoardo Arnaudo and Antonio Tavera contributed equally to this work.)

REFERENCES

[1] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.

[2] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[3] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet–A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[4] B. Chen, M. Xia, and J. Huang, "MFANet: A multi-level feature aggregation network for semantic segmentation of land cover," *Remote Sens.*, vol. 13, no. 4, p. 731, Feb. 2021.

[5] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.

[6] K. Nogueira, M. D. Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, "Learning to semantically segment high-resolution remote sensing images," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3566–3571.

[7] A. Farasin, L. Colomba, and P. Garza, "Double-step U-Net: A deep learning-based approach for the estimation of wildfire damage severity through Sentinel-2 satellite data," *Appl. Sci.*, vol. 10, no. 12, p. 4332, Jun. 2020.

[8] R. B. Andrade, G. A. O. P. Costa, G. L. A. Mota, M. X. Ortega, R. Q. Feitosa, P. J. Soto, and C. Heipke, "Evaluation of semantic segmentation methods for deforestation detection in the Amazon," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2020, pp. 1497–1505, Aug. 2020.

[9] M. T. Chiu, X. Xu, Y. Wei, Z. Huang, A. G. Schwing, R. Brunner, H. Khachatrian, H. Karapetyan, I. Dozier, G. Rose, D. Wilson, A. Tudor, N. Hovakimyan, T. S. Huang, and H. Shi, "Agriculture-vision: A large aerial image database for agricultural pattern analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2828–2838.

[10] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5367–5376, Aug. 2020.

[11] A. Tavera, E. Arnaudo, C. Masone, and B. Caputo, "Augmentation invariance and adaptive sampling in semantic segmentation of agricultural aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1656–1665.

[12] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, vol. 1, 2021.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[14] S. Chen, X. Jia, J. He, Y. Shi, and J. Liu, "Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11013–11022.

[15] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "DACS: Domain adaptation via cross-domain mixed sampling," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2021, pp. 1379–1389.

[16] L. Hoyer, D. Dai, and L. Van Gool, "DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9924–9935.

[17] Q. Zhou, Z. Feng, Q. Gu, J. Pang, G. Cheng, X. Lu, J. Shi, and L. Ma, "Context-aware mixup for domain adaptive semantic segmentation," 2021, *arXiv:2108.03557*.

[18] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "ClassMix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1369–1378.

[19] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Comp. Computer-Assisted Intervention*. Munich, Germany: Springer, 2015, pp. 234–241.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[26] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 34, 2021, pp. 9355–9366.

[27] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7262–7272.

[28] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.

[29] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2021.

[30] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12416–12425.

[31] J. Han, J. Ding, N. Xue, and G.-S. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2786–2795.

[32] E. Arnaudo, F. Cermelli, A. Tavera, C. Rossi, and B. Caputo, "A contrastive distillation approach for incremental semantic segmentation in aerial images," in *Image Analysis and Processing*. Cham, Switzerland: Springer, 2022, pp. 742–754.

[33] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and A. I. Ayed, "Boundary loss for highly unbalanced segmentation," in *Proc. Int. Conf. Med. Imag.*, vol. 102, 2019, pp. 285–296.

[34] B. Geng, D. Tao, and C. Xu, "DAML: Domain adaptation metric learning," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2980–2989, Oct. 2011.

[35] M. Long, Y. Cao, J. Wang, and I. M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn*, 2015, pp. 97–105.
[36] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," Dept. Comput. Sci., Princeton Univ., Princeton, NJ, USA, UC Berkeley, Berkeley, CA, USA, Tech. Rep., 2014.
[37] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.
[38] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2507–2516.
[39] H. Wang, T. Shen, W. Zhang, L. Duan, and T. Mei, "Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 642–659.
[40] A. Tavera, F. Cermelli, C. Masone, and B. Caputo, "Pixel-by-Pixel cross-domain alignment for few-shot semantic segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1626–1635.
[41] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.
[42] Z. Wu, X. Han, Y.-L. Lin, M. G. Uzunbas, T. Goldstein, S. N. Lim, and S. L. Davis, "DCAN: Dual channel-wise alignment networks for unsupervised scene adaptation," in *Proc. Int. Conf. Comput. Vis.*, 2018, pp. 518–534.
[43] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4085–4095.
[44] Q. Lian, L. Duan, F. Lv, and B. Gong, "Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6758–6767.
[45] Y. Zou, Z. Yu, B. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 289–305.
[46] K. Mei, C. Zhu, J. Zou, and S. Zhang, "Instance adaptive self-training for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 415–430.
[47] E. Alberti, A. Tavera, C. Masone, and B. Caputo, "IDDA: A large-scale multi-domain dataset for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5526–5533, Oct. 2020.
[48] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[49] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6936–6945.
[50] X. Wang, Y. Jin, M. Long, J. Wang, and M. I. Jordan, "Transferable normalization: Towards improving transferability of deep neural networks," in *Proc. Adv. Neural Inform. Process. Syst.*, 2019, pp. 1953–1963.

EDOARDO ARNAUDO received the M.S. degree in computer science with specialization in machine learning and artificial intelligence from the Università degli Studi di Torino (UniTo), in 2019, and the Ph.D. degree in computer vision from the Politecnico di Torino (Polito) in 2020, focusing on semantic segmentation applied to aerial and satellite imagery, which comprise his main research interests. In 2019, he started working at Links Foundation, Turin, Italy, as an Applied Researcher in the AI, data, and space domain in several multidisciplinary projects, with a focus on aerial and remote sensing applications.

ANTONIO TAVERA (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in computer science and software engineering from the Politecnico di Torino, Turin, Italy, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree in computer vision with the Visual and Multimodal Applied Learning Laboratory. He worked as a Master Thesis Intern at Italdesign Giugiaro S.p.A., in 2019. His research interests include semantic segmentation and domain adaptation applied to autonomous driving and aerial and satellite imagery.

CARLO MASONE (Member, IEEE) received the B.S. and M.S. degrees in control engineering from Sapienza University, Rome, Italy, in 2006 and 2010, respectively, and the Ph.D. degree in control engineering from the University of Stuttgart, Stuttgart, Germany, in collaboration with the Max Planck Institute for Biological Cybernetics (MPI-Kyb), Stuttgart, in 2014. From 2014 to 2017, he was a Postdoctoral Researcher at the Autonomous Robotics and Human-Machine Systems Group, MPI-Kyb. From 2017 to 2020, he worked in industry on the development of self-driving cars. From 2020 to 2022, he was a Senior Researcher at the Visual and Multimodal Applied Learning Laboratory, Politecnico di Torino, where he is currently an Assistant Professor.

FABRIZIO DOMINICI received the M.S. degree in telecommunications engineering from the Politecnico di Torino, in 2005. In 2005, he worked as a Researcher at the Istituto Superiore Mario Boella (ISMB), focusing on satellite and navigation systems. In 2013, he became the Head of research of the mobile solutions area at ISMB, now Links Foundation, and the Director of the Microsoft Innovation Center Torino (MIC), from 2012 to 2019. He currently manages and leads a large research group working in domains, such as IA, big data, GNSS technologies, and geospatial services. He is also the Head of Links Foundation, at the AI, data, and space research area. He has valuable experience managing multidisciplinary innovation projects at national and European levels. He supports the European Commission as an expert and he has been a coordinator of several EU-funded projects.

BARBARA CAPUTO received the Ph.D. degree in computer science from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2005. From 2007 to 2013, she was a Senior Researcher at Idiap-EPFL. Then, she moved to Sapienza Rome University, thanks to a MUR professorship, and joined the Politecnico di Torino, in 2018. She is currently a Full Professor with the Politecnico di Torino, where she leads the Artificial Intelligence (AI) Hub, Politecnico di Torino. Since 2017, she has been a double affiliation with the Italian Institute of Technology (IIT). She is one of the 30 experts who contributed to write the Italian strategy on AI and coordinator of the Italian National Ph.D. on AI and industry 4.0, sponsored by MUR. She is an ERC Laureate, ELLIS Fellow, and since 2019, she serves on the ELLIS Board.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement