Unsupervised Domain Adaptation for Instance Segmentation: Extracting Dwellings in Temporary Settlements across Various Geographical Settings

Getachew Workineh Gella, *Student member, IEEE*, Charlotte Pelletier, Sébastien Lefèvre, *Senior Member, IEEE*, Lorenz Wendt, Dirk Tiede, *Member, IEEE*, Stefan Lang, *Member, IEEE*

***Abstract*—Dwelling information is essential for humanitarian emergency response during or in the aftermath of disasters, especially in temporary settlement areas hosting forcibly displaced people. To map dwellings, the integration of very high-resolution remotely sensed imagery in computer vision models plays a key role. However, state-of-the-art deep learning models have two known downsides: (1) lack of generalization across space and time under changing scenes and object characteristics, and (2) extensive demand for annotated samples for training and validation. Both could pose a critical challenge during an emergency. To bypass this problem, this study deals with unsupervised domain adaptation for instance segmentation using a single-stage instance segmentation model, namely segmenting objects by location (SOLO). The goal is to adapt a SOLO model trained on a labelled source domain to detect dwellings in an unlabelled target domain. In this context, we study three domain adaptation techniques based on adversarial learning, domain discrepancy, and domain alignment mapping. We also propose domain similarity at different levels to understand its implication on domain adaptation. Experiments are conducted on very high-resolution satellite images obtained from four temporary settlement areas located in different countries and exhibiting various spatial characteristics. Analysis results show that in most source-target combinations unsupervised domain adaptation improves the performance by a large margin even surpassing a model trained with supervised learning. There is also an observed performance deviation among implemented strategies and different source-target dataset combinations. From the in-depth analysis of domain similarity at the image, object, and deep feature space levels, the former is more correlated with unsupervised domain adaptation performance.**

***Index Terms*—dwelling extraction, humanitarian response, unsupervised domain adaptation, instance segmentation, deep learning, domain similarity**

I. Introduction

TEMPORARY settlements and shelters host a significant number of forcibly displaced people (FDP) worldwide. Given the need for relevant information to monitor FDP settlement areas and synchronize humanitarian emergency response, remote sensing technology provides spatially detailed and repetitive observations from space. The information obtained from remotely sensed imagery is traditionally used to map camp expansion [1], [2] and infrastructure development [3], estimate resident population [4], [5], or assess environmental situations in FDP settlements [2], [6] using various classification and rule-set approaches. Similarly, there are studies dedicated to dwelling extraction from remotely sensed imagery using object-based image analysis (OBIA) and rule-set approaches [7], [8], [9], [10], [11].

When frequent monitoring of a specific settlement is required or when the geographic setting changes, workflows based on manual digitization and knowledge-based rule sets using the OBIA approach were challenged to meet the response time required to generate relevant information. Coupled with the availability of very high-resolution satellite imagery and advances in computer vision, deep learning models are paving the way for the automatic building detection. Leveraging this development, current instance segmentation models are able to localize and segment individual object instances from 2D- [12] and 3D-image scenes [13]. Though these developments resulted in the generation of global building footprint datasets [14], [15], FDP settlements are still less represented in terms of geographic coverage and provision of information with detailed spatio-temporal granularity [16]. As a result, recent promising works have focused on dwelling extraction in temporary settlements [17], [18], [19], [20], [21], building extraction in complex urban settings for humanitarian applications [22], and FDP settlement densification and spatial dynamic analysis [1].

Despite the proven performance of deep learning models for classification, segmentation, and detection tasks, they also have known limitations including their demand for bulk annotated data for training and validation and lack of generalization in different data distributions caused by changing scene and object characteristics [23]. In operational emergency response scenarios, the generation of training annotations is time-consuming and sometimes impractical given the time pressure for immediate emergency response. In post-emergency, FDP settlements are expanding, and dwelling structures (shelters, tents, facility buildings, tukuls, etc.) are changing in terms of size, shape, and spectral characteristics either because of new establishments, natural morphology, interventions that change rooftops or seasonal changes. These temporal variations are bottlenecks for the temporal transferability of trained models, requiring the generation of annotations for each image for frequent monitoring. By the same token, across geography, such settlements exhibit heterogeneous dwelling structures

G.W. Gella, L. Wendt, S. Lang and D. Tiede are in Christian Doppler Laboratory for Geospatial and EO-Based Humanitarian Technologies (GEOHUM), Paris Lodron University of Salzburg (PLUS), Salzburg, Austria

C. Pelletier and S. Lefèvre are in the Environment Observation with Complex Imagery (OBELIX) research lab, University of South Brittany(UBS), Vannes, France

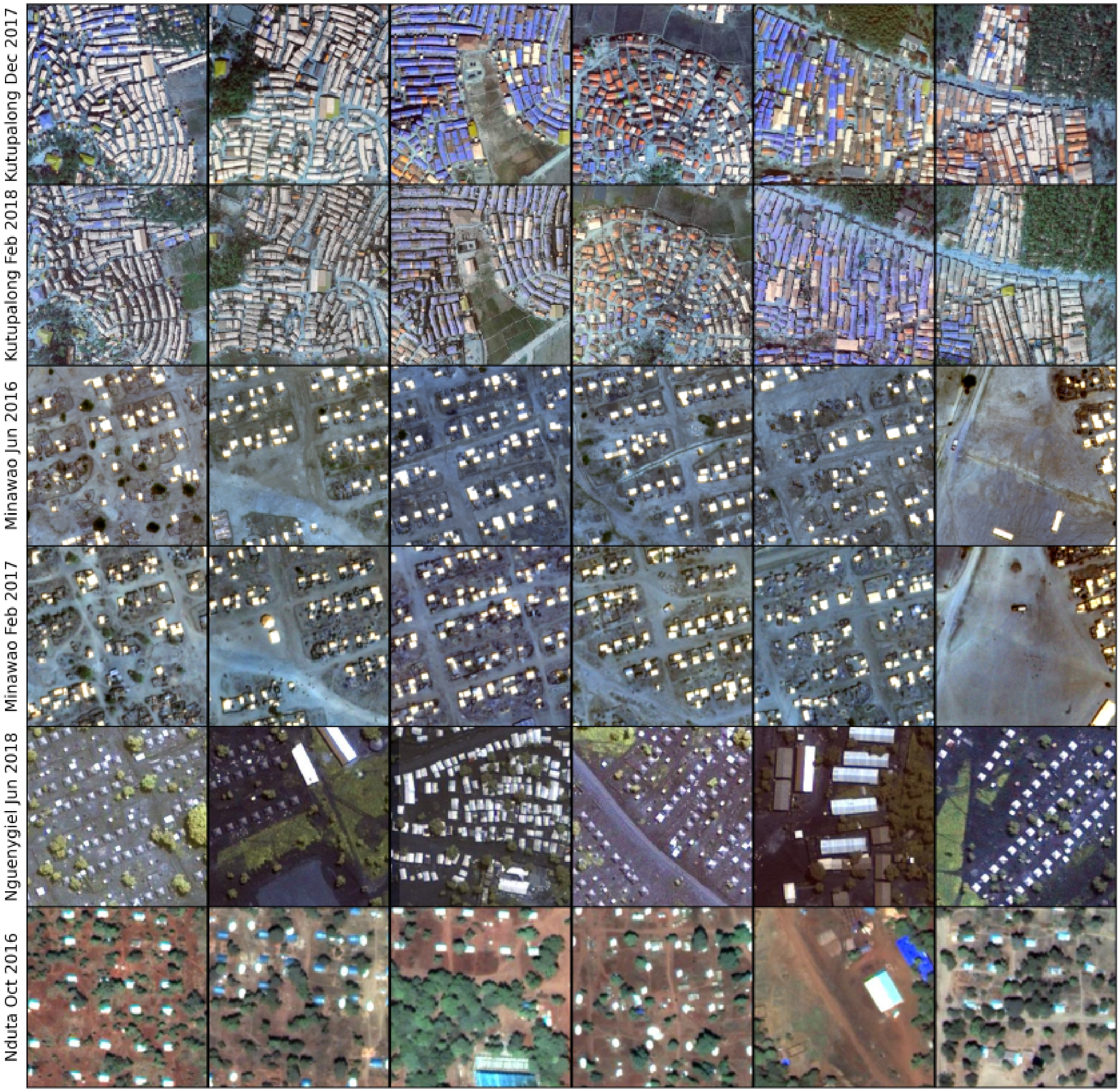

Fig. 1. Visualization of inter and intra-domain variations of dwelling spectral characteristics. For visual quality, images are scaled to 8-bit scaling.

and background environments. In this regard, Fig. 1, shows randomly selected images of different settlements. It highlights the inter- and intra-spectral variations of dwelling objects and background environment, which cause disparities in the corresponding feature space as it will be shown later in section II-B. This constrains the universal usage of models trained in one geography to perform similar tasks in datasets taken from different geographies without further retraining.

To overcome domain shifts caused by variations in a rural and urban setting [24], geography, season, sensor characteristics and sensor geometry, sensing domain, and inconsistencies in object classes [25], various transfer learning strategies have been devised. Transfer learning from pre-trained models with a large number of natural images [24], [26], [27], and fine-tuning of models trained with a large number of source data with a small labeled target set [28] are notable examples. These strategies still demand a significant amount of annotated data to retrain the model. Even with the availability of some samples, sometimes a finetuned model fails to effectively perform an intended task on the target datasets [29]. In circumstances where annotations do not exist during model training (which is also very common during emergency response operations), unsupervised domain adaptation strategies [30] could be viable alternatives. Unsupervised domain adaptation leverages labeled source data to learn representations enabling a model to undertake an intended task in the target domain without performance degradation. This is achieved by a joint training approach where representations are learned by optimizing a model using a combination of supervised and unsupervised losses. The model learns both semantically meaningful features on the main task for source dataset[31] and domain invariant features both for source and target datasets [32], [33] so that it can be applied to the target domain [30], [31].

Using remotely sensed imagery, unsupervised domain adaptation has been applied to (pixel and scene) classification and segmentation tasks. For example, hyperspectral image classification [34], [35], change detection [36], [37], cloud detection [38], land cover and scene classification [39], [40], [41], target detection and building extraction [42], [43] are recent notable examples for the application of unsupervised domain adaptation in earth observation datasets. As can be understood from an in-depth review that reveals advances in unsupervised domain adaptation with Earth observation (EO) imagery [44], [25], despite recent advances in unsupervised domain adaptation [45] in one way and instance segmentation in another way [12], [13], the combination of domain adaptation with instance segmentation for EO datasets is overlooked. Hence, the main emphasis of this article is to explore unsupervised domain adaptation approaches for dwelling extraction from very high-resolution (VHR) satellite imagery with a focus on EO-based humanitarian emergency response using a state-of-the-art single-stage instance segmentation model, namely Segmenting Objects by Location (SOLO) [46], [47]. Under this broader objective, the study has the following contributions:

1) Explored unsupervised domain adaptation for the instance segmentation to learn across space and time for dwelling extraction. This study was conducted using six very high-resolution EO images obtained from four FDP settlement areas.
2) Implemented and adapted three unsupervised domain adaptation strategies – Domain Adversarial Training of Neural Network (DANN), Maximum Mean Discrepancy (MMD), and Optimal Transport (OT) – for instance segmentation of dwellings from very high-resolution satellite imagery.
3) Conducted a comprehensive analysis of domain similarity using images, objects, and deep features, along with its implications for unsupervised domain adaptation transfer performance. This analysis will empower practitioners in operational emergency response settings to select a source dataset that ensures a positive transfer in the target dataset. A selected source dataset that has better domain similarity with the target dataset is expected to yield similar decision boundaries to the target task [48], facilitating a more effective transfer of learned representations.

Given these contributions, the remainder of this article is organized as follows. Section II details the methodology, while section III describes data and provides implementation details about the experimental setup. In section IV, results are presented and section V discusses obtained findings. Finally, we draw conclusions in section VI and give brief perspectives for follow-up work.

II. Method

In this section, we present the adopted methodology. We first provide a detailed description of unsupervised domain adaptation for instance segmentation. Then, we propose three strategies to compute domain similarity at different levels. Finally, we detail the evaluation metrics used for domain adaptation performance in section III-C.

A. Unsupervised domain adaptation for instance segmentation

As justified in section I, during an emergency scenario we assume the availability of both a source domain dataset D^s with a set of image and label pairs as $D^s = \{(\mathbf{x}_1^s, y_1^s), (\mathbf{x}_2^s, y_2^s), (\mathbf{x}_3^s, y_3^s), \ldots, (\mathbf{x}_n^s, y_n^s)\}$ and a target domain dataset D^t that contains a set of unlabeled images $D^t = \{\mathbf{x}_1^t, \mathbf{x}_2^t, \mathbf{x}_3^t, \ldots, \mathbf{x}_m^t\}$. Each image ($\mathbf{x}_i^s$ and $\mathbf{x}_i^t$) is a three-dimensional patch in $\mathbb{R}^{d\times h\times w}$, where h, w, and d correspond to the height, width and the number of spectral channels in the patch. In this study, each patch has a size of $3 \times 256 \times 256$. The label y_i^s represents a set of instance masks, each of dimension $h \times w$. Each individual instance mask is associated with a category. In this study, it represents the presence or absence of a dwelling.

The objective is accurate instance segmentation of the target domain without having access to target labels. Specifically, our aim is to learn representations that are both domain-invariant and discriminative to undertake instance segmentation of a target dataset by joint training from the labeled source and unlabelled target data [25]. To achieve this objective, we have implemented unsupervised domain adaptation strategies to serve the instance segmentation task. We used adversarial training [49] and domain discrepancy [33] approaches which both tried to address domain adaptation by focusing on deep feature space. We also explored a domain mapping approach [50], [51], [52], which tries to close the domain gap by considering both deep feature and label spaces.

1) Segmenting Objects by Location: Most unsupervised domain adaptation strategies can be adapted with less effort to different supervised tasks with any deep learning models [31], including instance segmentation. In this study, we used Segmenting Objects by Location (SOLO-v2) [47] as a base model, which is a single-stage instance segmentation model. The model is selected mainly for its reported speed, state-of-the-art performance, and relatively moderate memory requirement during training and inference phases [46], [47], which makes the model an ideal candidate for operational use during emergency response. Recently, the authors in [53] also demonstrated that on a few metrics, SOLO-v2 performs better than other models in urban building instance segmentation tasks. This was also verified in this study from a comparative analysis undertaken with other single and two-stage instance segmentation models using one of the datasets studied in this work, namely the Minawao June 2016 dataset presented in section III-A (see Table V in Appendix for results). To the best of our knowledge, this is the first use of the SOLO-v2 model for dwelling extraction on the one hand and unsupervised domain adaptation for instance segmentation on the other hand.

The model follows the proposal-free segmentation approach [54], where the input image is first conceptually divided into smaller tiles with different scales. The sizes of the objects to be segmented induce the scales to be taken into account. Scenes with small objects should be divided into many small tiles (grids), whereas scenes with large objects require a few larger tiles. In our application, the model must segment objects (i.e., dwellings) of varying sizes and thus consider different scales. In SOLO-v2, the grid size hyperparameter gives the set of scales to be studied. For our study, we selected its values by analyzing the sizes of the dwellings in the different study sites. Section III-B provides implementation details.

The SOLO-v2 model is composed of a feature extraction network, a Feature Pyramid Network (FPN) and a segmentation head. As indicated in Fig 2, an input image of size $3 \times 256 \times 256$ is fed into ResNet-50 [55] feature extractor, where it yields multilevel features. These multilevel features are fed into FPN [56] to get high-level semantic features of varying resolutions. To obtain precise locations of each object, in the deeper levels of FPN, SOLO-v2 introduced spatially invariant convolution using the CoordConv algorithm [57], where two extra channels containing the coordinates of each pixel are added. Please note that these pixel coordinates are relative row and column pixel coordinates (indexes), not absolute geographic coordinates. These multilevel features are fed into two sub-branches of the SOLO-v2 segmentation head, which are the unified mask branch and the kernel branch. These branches are decoupled for the sake of speed and memory optimization. The unified mask branch produces unified (i.e., fused) features of shape $256 \times 64 \times 64$ at a coarse spatial resolution (1/4 of the input image), whereas the kernel branch produces categories and kernel predictions at each grid level.

During the training phase, each ground truth instance mask is assigned to any tile for which the dwelling's centre falls within the tile. Otherwise, it is considered as background (negative tile) [47]. During inference, the predicted instance masks and their corresponding labels are obtained in a two-step procedure. First, the predicted category score is thresholded to determine a pool of candidate instances. Then, predicted instance masks are converted to binary instance masks using a mask threshold. Finally, further refinement on the overlapping instance masks is done using Matrix Non-Maximum Suppression (MNMS), which is also reported as computationally lighter than traditional NMS [46]. MNMS provides a modified category score using the decay function, which is hereafter renamed as the objectness score. The objectness score is a score indicating the degree a candidate mask contains an object of interest, i.e., dwelling. After sorting predicted instance masks with descending order of objections score, the first k number of final instances was picked where k is the allowed maximum number of detections per scene. As all the predictions are done at a coarse spatial resolution, the predicted instance masks are up-sampled to the original input size using bi-linear interpolation. In a nutshell, the inference provides predicted object instance masks with the corresponding categories (a semantic class) and objectness scores.

During training, the model weights are optimized by minimizing a combination of instance mask and category losses presented in Eq. 1 as:

$$\mathcal{L}^S = \mathcal{L}^C + \lambda \mathcal{L}^M \tag{1}$$

where $\mathcal{L}^S$ is a total supervised loss, $\mathcal{L}^C$ is a category loss (here the Focal loss [58]), $\mathcal{L}^M$ is the instance mask loss (here the Dice loss [59]), and λ is the weighting factor for balancing the contribution of each loss term. Implementation details of SOLO-v2 are provided in [46], [47].

2) Unsupervised domain adaptation: This section presents the details about implemented unsupervised domain adaptation strategies used to conduct the different experiments. Though there are many unsupervised domain adaptation approaches, we opted for three representative approaches that fall under the broader category of unsupervised deep domain distribution alignment approaches [30]. It should be noted that our specific choice for this broader category of domain adaptation is based on two assumptions. The first is that the intended task (i.e., instance segmentation) focuses on the same class of interest in the two domains (dwellings versus background). There is thus only a distribution shift due to the changes in scene and object characteristics. The second is the absence of annotations for the target domain. Hence, the objective is to use domain adaptation strategies to reduce the distribution shift, and thus be able to perform instance segmentation on the unlabeled target dataset.

Therefore, from the stated broader category, we have opted to implement DANN, MMD, and OT approaches. Our specific choice of these strategies is mainly because (1) DANN and MMD are standard unsupervised domain adaptation approaches that demonstrate good results in various experiments (classification and segmentation) so seem appropriate candidates to experiment for an instance segmentation task, while OT has shown good performance in computer vision tasks; and (2) the strategy followed by these three approaches to address the unsupervised domain adaptation problem differ as it will be explained in the next paragraphs. This could give us a chance to select the best-performing domain adaptation strategy for operational use in the humanitarian emergency response.

The first unsupervised domain adaptation approach tested is Domain Adversarial Training of Neural Networks (DANN) [49]. DANN adversarially trains the feature extractor to bring source and target data distributions into a relatively common feature space. This is performed by implementing a domain classifier module, whose goal is to discriminate the domain (source or target) from which data is sampled. The feature encoder parameters are updated to minimise the supervised loss $\mathcal{L}^S$ computed on the source data (Eq. 1) and to maximise the adversarial domain classifier loss $\mathcal{L}^{AD}$ (negative log-likelihood). This adversarial learning procedure is ensured by the gradient reversal layer [49], which multiplies the gradients from the domain classifier layer by a negative constant during the back-propagation. The overall loss

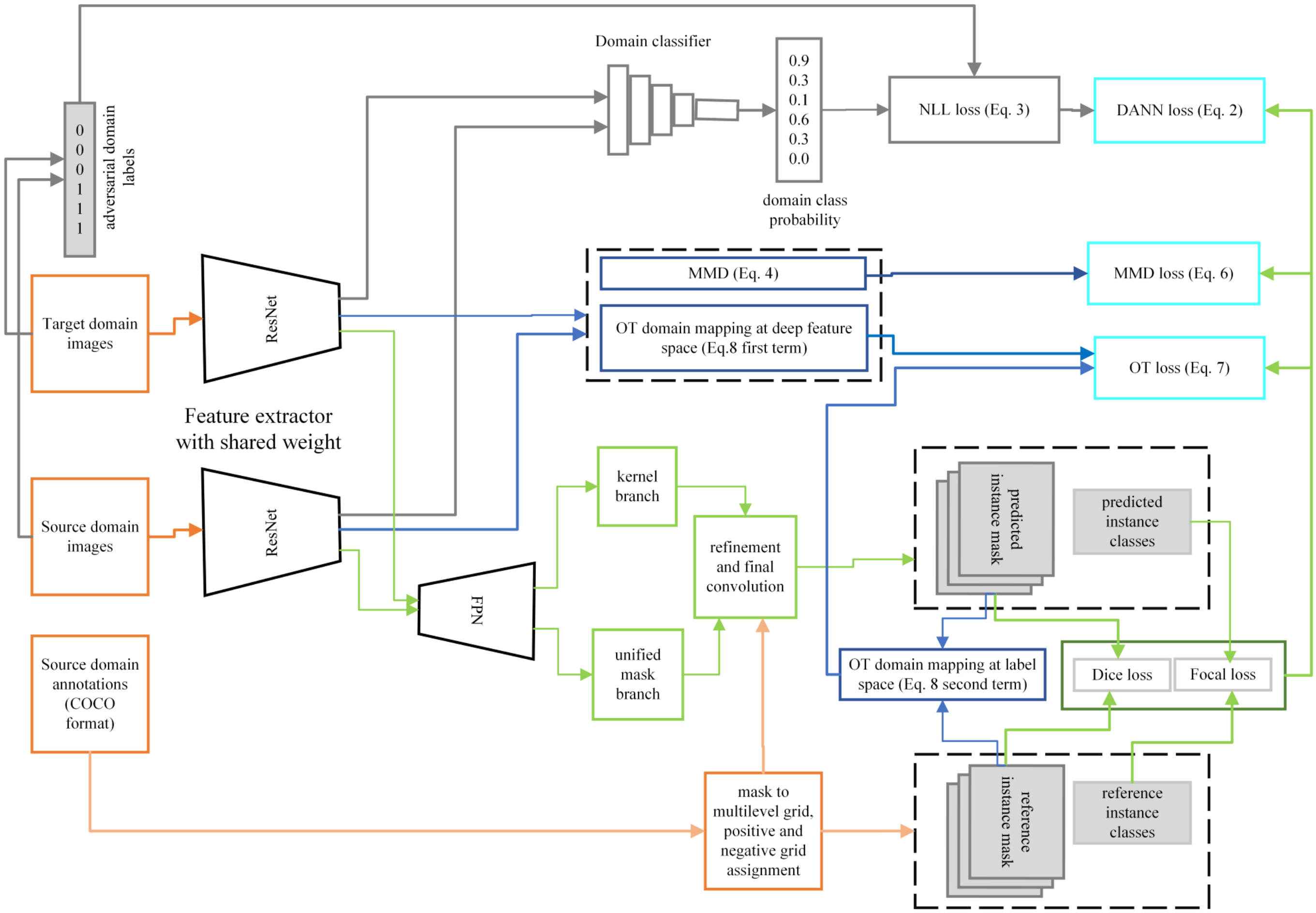
Domain classifier
0.9
0.3
0.1
0.6
0.3
0.0
domain class probability
NLL loss (Eq. 3)
DANN loss (Eq. 2)
0
0
0
1
1
1
adversarial domain labels
Target domain images
ResNet
Feature extractor with shared weight
Source domain images
ResNet
MMD (Eq. 4)
OT domain mapping at deep feature space (Eq.8 first term)
MMD loss (Eq. 6)
OT loss (Eq. 7)
kernel branch
refinement and final convolution
FPN
unified mask branch
predicted instance mask
predicted instance classes
OT domain mapping at label space (Eq. 8 second term)
Dice loss
Focal loss
Source domain annotations (COCO format)
mask to multilevel grid, positive and negative grid assignment
reference instance mask
reference instance classes

Fig. 2. Schematic representation of overall workflows for unsupervised domain adaptation using SOLO-v2. Elements indicated with orange color are inputs and ResNet is a shared feature extraction network used for both source and target datasets. Green lines indicate the supervised pipeline for the source dataset where features from shared feature extractor are fed into the segmentation network, for which category (Dice) and mask (Focal) losses are calculated. For Domain Adversarial Training of Neural Network (DANN) (grey arrows), features from ResNet are concatenated batch-wise and fed into a domain classifier where negative log-likelihood (NNL) loss is computed using adversarial domain labels. Blue lines indicate unsupervised loss computed from feature and label spaces both for Maximum Mean Discrepancy and Optimal Transport unsupervised loss terms.

$\mathcal{L}^{DANN}$, provided in Eq. 2, consists of both supervised and domain losses:

$$\mathcal{L}^{DANN} = \mathcal{L}^{S} + \lambda_u \mathcal{L}^{AD} \tag{2}$$

where λ_u is the weighting factor for the contribution of the unsupervised loss. The unsupervised adversarial domain loss $\mathcal{L}^{AD}$ is the Negative Log-likelihood (NLL) loss which is provided as:

$$\mathcal{L}^{AD} = -\sum_{i=1}^{n+m} (d_i \log \hat{d}_i + (1-d_i)\log(1-\hat{d}_i)) \tag{3}$$

where d_i is adversarial domain label (d_i is equal to 1 if the i-th sample belongs to the target domain, 0 otherwise), while $\hat{d}_i$ is predicted domain label, n and m are the number of samples for the source and target datasets, respectively.

The second domain adaptation approach tested is Maximum Mean Discrepancy (MMD) [33]. It adds an unsupervised loss, which accounts for inter and intra-domain discrepancy in the deep feature space. Specifically, the domain discrepancy loss $\mathcal{L}^{DANN}$ measures the domain discrepancy between the source (S) and target (T) deep features using MMD [60], [61]. This domain discrepancy loss $\mathcal{L}^{DD}$ can be expressed as

$$\begin{aligned}\mathcal{L}^{DD}(S,T) = &\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j=i}^{m} k(g(\mathbf{x}_i^s), g(\mathbf{x}_j^s))\\ &+\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i}^{n} k(g(\mathbf{x}_i^t), g(\mathbf{x}_j^t))\\ &-\frac{2}{(mn)}\sum_{i=1}^{m}\sum_{j=i}^{n} k(g(\mathbf{x}_i^s), g(\mathbf{x}_j^t))\end{aligned} \tag{4}$$

where $g(.)$ is the feature extractor network (ResNet-50), and $g(\mathbf{x}_i^s)$ (resp. $g(\mathbf{x}_i^t)$) represents the embedding of the i^{th}-sample in the source (resp. target) dataset, k is a Gaussian kernel (defined by Eq. 5), and n (resp. m) is the number of source (resp. target) samples:

$$k(g(\mathbf{x}_i^s), g(\mathbf{x}_j^t)) = e^{\frac{-|g(\mathbf{x}_i^s)-g(\mathbf{x}_j^t)|^2}{2\delta^2}} \tag{5}$$

where δ is the standard deviation. Finally, the total loss $\mathcal{L}^{MMD}$ comprises a joint supervised loss $\mathcal{L}^{S}$ computed by the instance segmentation network and the domain discrepancy loss $\mathcal{L}^{DD}$ as provided in Eq. 6:

$$\mathcal{L}^{MMD} = \mathcal{L}^{S} + \lambda_u \mathcal{L}^{DD} \tag{6}$$

The third and last domain adaptation approach relies on the optimal transport theory, specifically Deep Joint Distribution Optimal Transport (DeepJDOT) [50], [51]. Similar to DANN and MMD, the overall loss $\mathcal{L}^{OT}$ is a combined loss including both supervised and unsupervised optimal transport terms:

$$\mathcal{L}^{OT} = \mathcal{L}^{S} + \lambda_u \mathcal{L}^{UOT} \tag{7}$$

where $\mathcal{L}^{S}$ is the supervised loss provided in Eq. 1 and $\mathcal{L}^{UOT}$ is the unsupervised optimal transport loss. Unlike DANN or MMD which focus only on obtaining domain-invariant features, the optimal transport loss also benefits from aligning the label space. It combines two unsupervised losses computed on the deep feature space and the label space:

$$\mathcal{L}^{UOT} = \sum_{i,j} \gamma_{i,j}(\alpha||g(\mathbf{x}_i^s) - g(\mathbf{x}_j^t)||^2 + \beta||y_i^s - f(g(\mathbf{x}_j^t))||^2) \tag{8}$$

where $g(\cdot)$ is the feature extractor network as in Eq. 4 and $f(\cdot)$ is the instance segmentation network. The hyperparameters α and β are the weighting factors for the feature- and label-space losses, respectively. As the study deals with instance segmentation task, elements in the second term of Eq. 8, y_i^s and $f(g(\mathbf{x}_j^t))$ indicates reference and predicted dwelling instances, respectively. The $\gamma_{(i,j)}$ is the optimal transport coupling matrix [52], which is the key element for optimal transport. It gives a mapping between source and target data. Its computation is done using the Python Optimal Transport (POT) package [62] with default values. We note $f(g(\mathbf{x}))$ corresponds to the prediction of the SOLO-v2 network, which commonly includes an instance mask and a semantic category for each detected instance. In our experiments, as there is only one semantic class (dwelling) we can thus simplify the label space loss by taking into account only the instance mask predictions.

B. Domain similarity measures

To understand the implication of domain similarity on the performance of unsupervised domain adaptation, the cross-domain similarity is computed at three levels. These are object level (size, spacing, and shape complexity), image level (spectral, radiometric, and tonal variations like texture), and deep feature space levels. In operational humanitarian response, this information could serve as a catalogue for proper source dataset selection to undertake joint training and testing between source and target datasets. Please note that compared to image and deep feature space levels, which do not require annotations, the object-level comparison requires supervision in both domains. We still assume, for images with good visual quality, domain experts in earth observation could easily make some interpretations on object similarities.

For object-level domain similarity, following the works [63], selected landscape metrics that are suitable to catch dwelling object geometric and spatial pattern variations across space are considered. Accordingly, we computed dwelling density as the number of dwellings per hectare, minimum distance to the nearest dwelling, and shape index, which is the ratio of perimeter per dwelling object area across domains. Details on the conceptual definitions and implementation of these metrics are provided in [64].

For domain similarity at the deep feature space level, following the works of [65], we use both visual (qualitative) and metric similarities at the feature space level. Accordingly, the visual analysis and the computation of the similarity metric of deep features are done using t-distributed Stochastic Neighbour Embedding (t-SNE) [66]. To this end, the deep features of each input image are obtained using a ResNet pre-trained with ImageNet dataset. We opted to use pre-trained weights because fine-tuning the network on actual data demands annotations both in source and target datasets. In the unsupervised domain adaptation setting, we assume the non-existence of annotations for the target dataset. The deep features of each input image were hooked from the feature extraction module. Then, their dimensionality is reduced to two components by setting the principal component as the initial embeddings given its reported better stability than random initialization. The fitting of dimensionality reduction in t-SNE is done using a perplexity value of 50 with 5000 iterations.

Please note that as there is no single standard approach for optimizing perplexity [67], we set this value based on trial and error with a visual comparison of obtained clustered feature spaces. As shown in section IV-B, outputs from t-SNE are visualized in a 2-D space as a qualitative understanding of domain similarity at the feature space level. The deep embeddings were also fitted using a One-class Support Vector Machine (OC-SVM) [68], to draw a decision boundary that separates inliers and outliers from provided feature vectors. The implementation of both t-SNE and OC-SVM is forked from [69]. Following the work of [65], Intersection over Union (IoU) between two inlier masks can be computed to quantify feature space similarity. However, IoU is limited if the OC-SVM produced non-overlapping inlier masks (i.e., IoU is null). Hence, we opted for the Generalized Intersection over Union (GIoU) [70] metric, which uses the distance between the inlier masks. It could provide feature space distance including for non-overlapping masks. Its value ranges from -1 for completely not overlapping features with varying distances and +1 for perfectly overlapping features. The GIoU formulation is given by:

$$GIoU = \frac{|M_s \cap M_t|}{|M_s \cup M_t|} - \frac{|H(M_s \cup M_t) \setminus (M_s \cup M_t)|}{|H(M_s \cup M_t)|} \tag{9}$$

where M_s and M_t are the source and target domain inlier masks respectively, and H the convex hull function. The proposed methodology is further described in [70].

Finally, image-level domain similarity is quantified using the Structural Similarity Index (SSIM) [71]. It accounts for the luminance, contrast, and structure of images using local and global statistics. The SSIM is defined by Eq. 10:

$$
\begin{aligned}
SSIM(S,T) &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} SSIM(\mathbf{x}_i^s, \mathbf{x}_j^t), \\
SSIM(\mathbf{x}_i^s, \mathbf{x}_j^t) &= \frac{(2\mu_s\mu_t + C_1)(2\sigma_{ts} + C_2)}{(\mu_s^2 + \mu_t^2 + C_1)(\sigma_s^2 + \sigma_t^2 + C_2)}
\end{aligned}
\qquad (10)
$$

where x_i^s (resp. x_j^t) is the i^{th} (resp. j^{th}) image sample in the source (resp. target) dataset, μ, σ and σ^2 indicate the mean, standard deviation, and variances of the source (s) and target (t) images within a given window size (11 × 11 pixels) and C_1, C_2 are constants used for numeric stability which are equal to 0.01 and 0.03, respectively. Details are provided in [71] and its base implementation is forked from [72]. The higher the SSIM, the higher the similarity of the images. A value of +1 (or -1) corresponds to perfectly (dis)similar images.

III. Data and experimental setup

This section presents details about the data and applied pre-processing operations with overall experimental setups.

A. Data and pre-processing

The study used multi-source, multi-temporal very high resolution (VHR) satellite imagery sensed from four FDP settlement areas situated in different geographical areas – Kutupalong, Minawao, Nduta, and Nguenygiel. Figure 3 shows the geographic location of the study sites, whereas Table I gives further information on the type of sensor, the acquisition date, and the number of annotations available.

TABLE I
Characteristics of Earth Observation datasets used for the study.

Study site	Country	Acq. date	Sensor	No. annotations
Kutupalong	Bangladesh	24/12/2017	UAV	110,038
Kutupalong	Bangladesh	13/02/2018	UAV	125,409
Minawao	Cameroon	03/06/2016	WorldView-3	16,083
Minawao	Cameroon	12/02/2017	WorldView-3	20,257
Nguenygiel	Ethiopia	22/06/2018	Pléiades-1	16,097
Nduta	Tanzania	21/10/2016	WorldView-2	10,631

Note: All images are resampled to 0.5-meter spatial resolution.

The selected camps exhibit contrasting spectral, object, and background characteristics. Minawao is located in a semi-arid climate and consists of oval and round-shaped dwellings built by the United Nations Higher Commission for Refugees (UNHCR). This FDP area is dominated by small tukuls with roof material made from natural leaves, resulting in a low contrast with the ground. Sharing object type and size variations (see Table II) with Minawao, Nduta has well-vegetated green individual trees which are posing a major confusion with detached buildings during inference. Contrary to these two study sites, the dwellings in Kutupalong occupy a complex and hilly terrain with extreme spectral and object geometric variations. More importantly, in some regions of the site, the dwelling density is so high that individual buildings cannot be separated, even during the visual interpretation.

TABLE II
Descriptive statistics for dwelling object sizes.

Dataset	Min	Q1	Q2	Mean	Q3	Max	Sk
Kutupalong-2017	2.0	11.0	18.0	24.8	29.0	740.0	4.5
Kutupalong-2018	2.0	10.6	18.4	25.2	29.7	742.6	4.8
Minawao-2016	2.0	6.6	11.8	12.9	14.5	973.2	19.4
Minawao-2017	2.0	7.0	12.0	13.4	15.0	712.0	13.8
Nduta-2016	2.0	5.6	17.7	20.4	28.8	1027.7	11.9
Nguenygiel-2018	5.0	15.5	22.8	25.6	30.0	495.6	9.2

Note: Q1, Q2, Q3 and Sk stand for first, second and third quartiles and Skewness values, respectively. All values are provided in square meters (m^2).

The images are obtained from different optical sensors. Respective annotations were obtained from an in-house database [73] generated during a long-term engagement in EO-based humanitarian emergency response. WorldView-2 and WorldView-3 images were obtained in ortho-ready standard GeoTIFF format, radiometrically corrected, and projected to the standard plane [74]. Terrain flattening was made using the Surface Radar Topographic Mission Digital Elevation Model (SRTM DEM). For both Pléiades and WorldView images, multi-spectral and panchromatic bands were merged using Gram-Schmidt pan-sharpening [75]. Unmanned Aerial Vehicle (UAV) images with RGB channels were taken by a Sensefly Photogrammetry Sensor Optimized for Drone Applications (SODA) sensor mounted on the eBee drone series. These images were provided by the International Organization for Migration (IOM) and shared at Open Aerial Map (OAM). For a fair transfer performance comparison and to reduce the complexity that comes from sensor variations, all images with different radiometric resolutions were scaled to an 8-bit dynamic range, enhanced with histogram stretch, and re-sampled to similar pixel sizes. After the pre-processing of images, manually annotated labels in Environmental Systems Research (ESRI) shapefile format were converted to binary (dwelling and background) raster image coregistered with corresponding spectral images. Beyond testing the domain adaptation performance of instance segmentation across different geographical areas, we analyzed the performance of domain adaptation across time for two study sites (Kutupalong and Minawao) that included additional images from the same area acquired at different years and seasons. To use the same model and run a fair comparison with UAV images, only RGB channels are used for all the satellite datasets.

The images and the related rasterized annotations are divided into smaller image patches with a size of 256 × 256 pixels each. The size is intended to accommodate a sufficient number of objects per patch without compromising computational speed. The individual patches are partitioned into train, test, and validation sets with a ratio of 70%, 15%, and 15%, respectively. During the patch partitioning, to account for the existing spatial heterogeneity of dwellings in each studied site, patches are selected by block sampling using their geolocation (geographic extent) as a constraint (see Fig. 4). The camp extent is divided into larger regular blocks where samples were picked using systematic sampling. Each block contains a maximum of 16 image patches. This sampling approach is mainly intended to balance two problems: the first one is

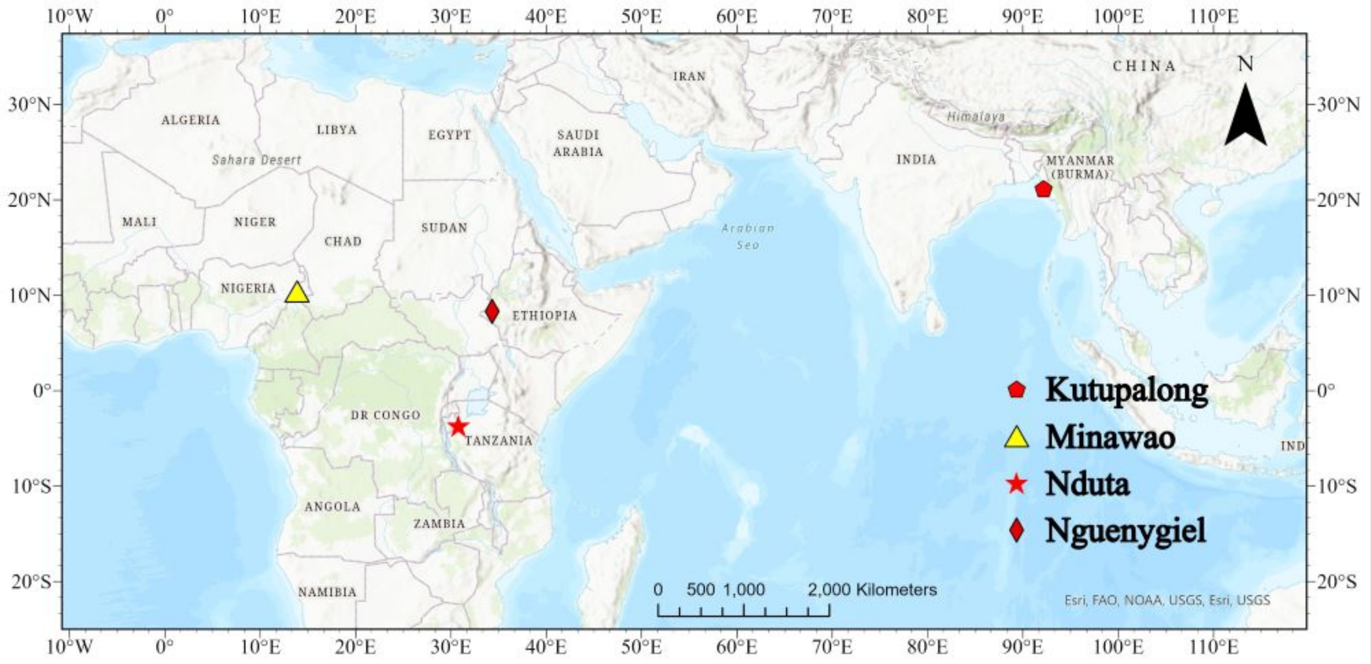

Fig. 3. Location map of temporary settlement areas in different geographies. Please note that for the purpose of visual quality, the point markers are not to scale

model over-fitting caused by the inclusion of mostly similar objects from adjacent tiles (spatial auto-correlation); and the second one is the reduction of out-of-distribution samples caused if the image is simply divided into 3 regions where each serves training, validation, and test sets, as implemented in [20]. Then, sampled patches and raster annotations are converted to the Common Objects Context (COCO) [76] format, which is suitable for the SOLO-v2 input-output structure. As can be seen from Fig. 4, empty tiles that did not contain dwelling objects are excluded from all sample sets to reduce the negative effects of class imbalance during the training phase.

B. Experimental setup

As shown in Fig. 4, the samples from each dataset are partitioned into training, validation, and test sets. The base model (here a SOLO-v2 with a ResNet-50 feature extractor) is trained with unsupervised domain adaptation approaches using labeled source data from the training and validation sets and unlabeled target dataset taken only from the training set. The performance of the different strategies is evaluated on the labeled target test set. We test exhaustively all the possible cross-domain adaptation scenarios. As there are six datasets, it results in 30 scenarios. Please note that we preferred to conduct an extensive study without considering any temporal constraint, which can lead to implausible scenarios (e.g., the Minawao image acquired in 2017 is used as the source domain and the 2016 image is used as the target domain). To better understand the performance of these unsupervised domain adaptation strategies for dwelling extraction, two baselines were undertaken. The first baseline is trained and tested with a fully-supervised approach on each dataset. It serves as an upper bound, *i.e.*, the maximum accuracy to reach. The second baseline is the lower bound, which corresponds to the non-adaptation case where the source-trained model is applied to the target domain without any adaptation.

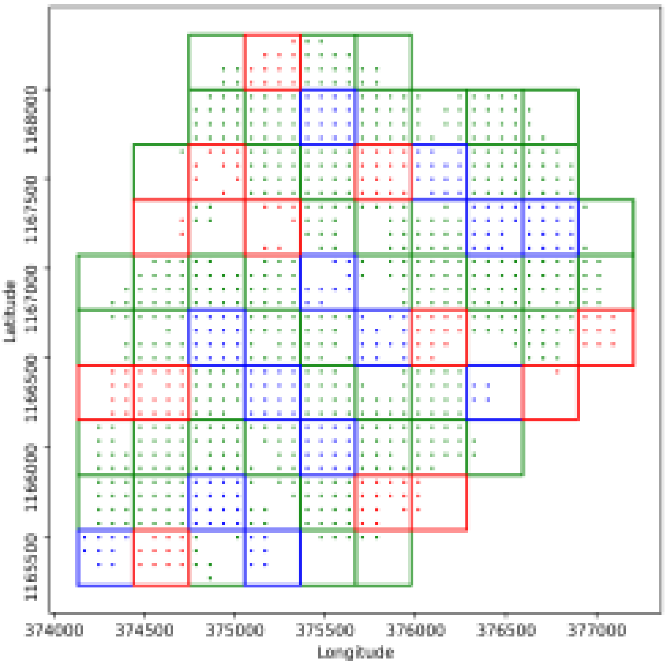

Fig. 4. Block-sample strategy scheme implemented to partition image patches into training (green), validation (blue), and test (red) datasets. The image is partitioned into large blocks that consist of 16 image patches, each with a size of 256 × 256 pixels represented by coloured dots. The example is given for the Minawao-2016 dataset.

Before running the domain-adaptation experiment, the instance segmentation model SOLO-v2 is first pre-trained on the source domain. To set up the main grid size hyperparameter, which determines the size of the detected objects by SOLO-v2 (see Section II-A1), we analyse the dwelling sizes. We decided to use a list of grid scales equal to 6, 12, 24 and 48, which represent the varying dwelling sizes for the different studied sites. The λ hyperparameter (Eq. 1), is used to balance the instance and semantic losses. Values ranging from 0.1 to 10 are tested, and the best test performance is obtained for λ equal to 4.

During the inference phase, predicted non-dwelling object candidates were first filtered out using a score threshold which is set to 0.3. The Matrix Non-Maximum Suppression [46], [47] is applied, which provides a modified category score as a function of decay value [46]. Then all predictions are sorted in descending order of modified score threshold, where only k number of detections were picked. k is the maximum number of detections per image which was set to 200 dwellings. This value is determined by accounting for the dwelling prevalence statistics per image patch. After the final inference, if there need to refine prediction results after qualitative inspection, the category score after MNMS (objectness score) (see II-A1) could also serve as a post-processing tool to remove over-predicted dwellings that have small objectness score. This could apply even for scenes with object predictions of less than 200 dwellings (e.g. see Fig. 12).

In the unsupervised domain adaptation setting, we do not have access to the annotated target dataset. Hence, the tuning of the value of the hyperparameter λ_u through cross-validation is not possible. We thus use a logarithmic annealing schedule (Eq. 11) where in the early phases of the training the contribution from unsupervised terms is given a lower weight than at the later stages where equal contribution with source domain loss is given. This creates a smooth training process by gradually learning from the unlabelled target data.

$$\lambda_u = \frac{2}{1 + e^{-\gamma p}} - 1 \tag{11}$$

where p corresponds to the normalized training step which ranges from 0 to 1, γ is a constant to determine the initial minimum value for λ_u which is set to 10 [49]. Regarding the specific hyperparameter of the optimal transport loss (see Eq. 8), we assume an equal contribution of deep feature space and label space terms, and thus set up α and β equal to 1.

In all the experiments (pre-training and domain adaptation), the network is updated with Stochastic Gradient Update (SGD) for 65 epochs. As in the original implementation of SOLO-v2, the multi-step learning rate decay is adopted with two change points where the initial learning rate is set to 0.001 and decreased by a gamma value of 10 of the previous steps. The training is done on a computer equipped with NVIDIA GeForce RTX 3090 single GPU.

C. Evaluation metrics

The performance of each cross-domain scenario is quantitatively evaluated using mean average precision (mAP) which is a commonly used metric in object detection and instance segmentation tasks [76], [77]. It is computed using the following equation:

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \tag{12}$$

where AP_c is the average precision of class c and C is the total number of classes. As there is only one class to predict, the dwelling class, the mAP corresponds in this study to the average precision of the dwelling class. To determine if a segmented object matches a ground-truth object, a mean Intersection over Union (mIoU) threshold of 0.5 is employed (mAP@0.5). This is used to compute precision and recall metrics. To understand the association between transfer performance and domain similarity, a statistical test is done using the Pearson correlation coefficient with 90% and 95% confidence intervals. To perform this test, image and object level metrics are summarized using mean and mean absolute differences across each transfer respectively.

IV. Results

In this section, we report results obtained from all unsupervised domain adaptation approaches presented in section II-A. In addition, analytical results of domain similarity at the image, object and deep feature space levels and its implication for unsupervised domain adaptation performance are presented in section IV-B.

A. Domain adaptation and transfer performance

Fig. 5 displays the mAP@0.5 for the 30 cross-domain transfer scenarios. Each subplot presents domain transfer made from a source dataset (x-ticks label) to a target dataset (which is indicated on the title of each subplot). The black dot lines correspond to the performance obtained without adaptation (lower bounds), whereas the red dot lines highlight the performance obtained by a fully supervised model trained on the target domain (upper bound). The bar corresponds to the performance obtained by the UDA approaches: MMD in blue, DANN in orange, and OT in green. domain adaptation has performed better than applying a source pre-trained model without domain adaptation. As it can be seen, in a few source-target combinations, domain adaptation performed even better than supervised training (for example in the following transfers: Nguenygiel-2018 → Minawao-2016, Nguenygiel-2018 → Minawao-2017, Nguenygiel-2018 → Nduta-2016). Among the three approaches, the OT-based approach has shown better performance than DANN and MMD approaches, which have comparable performance. Sometimes, DANN and MMD results are even lower than the performance of the source model applied without adaptation such as for transfers Kutupalong-2018 → Nduta-2016, Kutupalong-2017 → Kutupalong-2018, and Kutupalong-2018 → Minawao-2017. As it will be discussed in the next section, this is particularly noticeable for transferring between domains that have large disparities in spectral and object characteristics (for example transferring from Kutupalong-2018 to Minawao, Nguenygiel

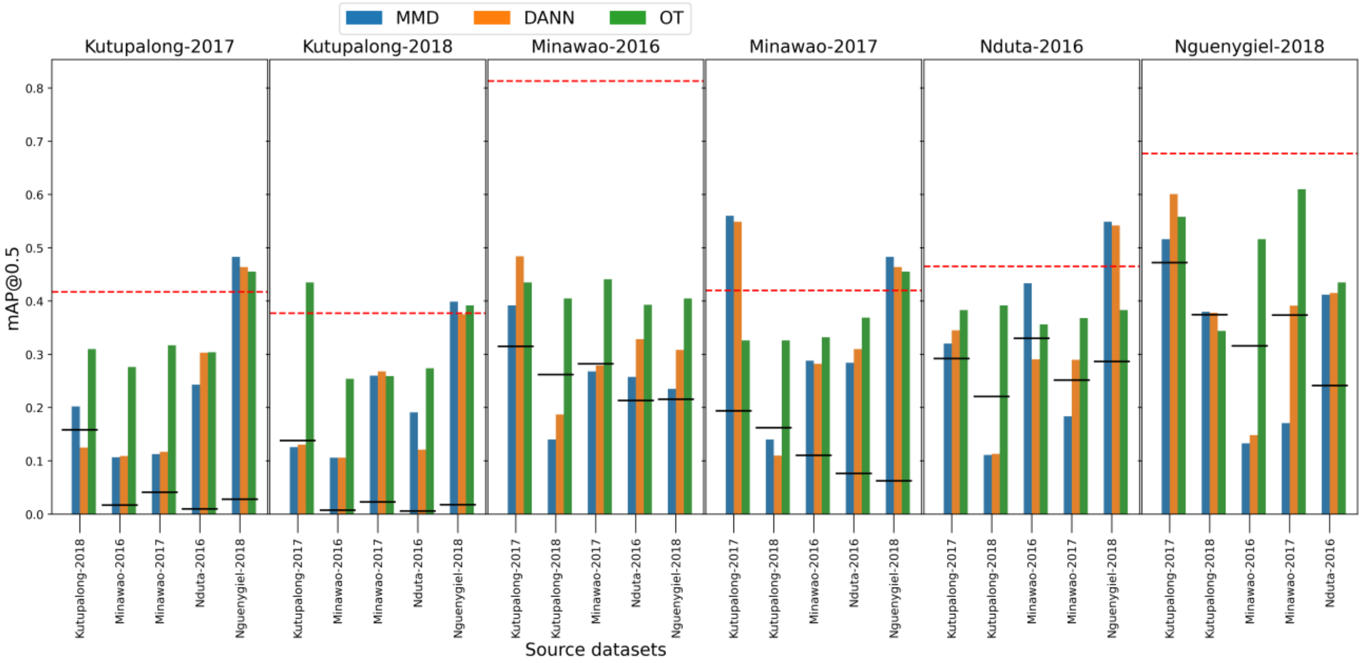

Fig. 5. Results for unsupervised domain adaptation for instance segmentation ($mAP@0.5$) using SOLO for dwelling extraction in FDP settlement areas. The title of each sub-plot indicates the target dataset, whereas the x-tick label indicates the source dataset. The red dotted horizontal lines shows the performance obtained with a fully supervised learning strategy, while the dotted black lines indicate the performance obtained by the source network applied without domain adaptation.

and Nduta datasets). The performance deviates from a mAP values ranging from below 0.1 up to approximately 0.6.

We note that the transfer performance decreases when transfer is done from a dataset with less complex dwelling objects dominated by standard UNHCR tents to the dataset dominated with most complex dwelling structures with densely populated and exhibit larger object spectral and size variations. This is observed for transfers made from Minawao, Nguenygiel, and Nduta to Kutupalong datasets. Contrary to this, transferring from datasets with more complex dwelling structures towards less complex dwelling structures provides better results than transferring from less complex to more complex sites. This could be attested by transfer results from Kutupalong-2017 to Minawao, Nguenygiel and Ndutda.

As can be seen from predicted spatial plots displayed in Fig. 6, transfers made without domain adaptation have relatively higher false negative rates since existing dwelling objects were not segmented, especially in transfers Kutupalong-2017 → Minawao-2016 and Kutupalong-2017 → Kutupalong-2018 (see Fig. 7 for visible insets). DANN has shown relatively higher over-predictions and false positives posed by individually standing trees and bare land features resembling low-contrast dwellings. These errors could easily be managed during the post-processing phase.

B. Transfer performance as a function of domain similarity

As noted in Section I, the main objective of undertaking domain similarity analysis is to establish a clue which similarity-level (image, object, or deep feature) is the most suitable to explain domain adaptation performance. This could further be used as a lookup table to select source datasets that could be easily transferred to the intended target dataset during humanitarian emergency response. Accordingly, we compare domain similarities at three levels based on images, objects, and deep features.

First, we look into image-level domain similarity by computing the SSIM metric provided in Eq. 10. As can be seen from computed similarities presented in Fig. 9, datasets show low to moderate image-level SSIM values (as indicated in Section II-B, the value of 1 means perfect similarity). Here we can observe three patterns of structural similarity and transfer performance. The first pattern is where a lower level of structural similarity is associated with a lower transfer performance (Fig. 6). One notable example is MMD and DANN-based transfers where Kutupalong-2018 is a source dataset (with the exception of OT-based transfers). With the same token, the second pattern is where a higher structural similarity is associated with higher performance. This could be easily noted between SSIM and transfer performances from Nguenygiel-2018 to Minawao (both 2016 and 2017) datasets. The third pattern is high transfer performance and relatively moderate structural similarity. This is vividly seen from transfers made between Kutupalong-2017 and the rest of the datasets except the Kutupalong-2018 dataset for OT-based transfer. Though Kutupalong-2017 is a dataset with relatively complex dwelling structures, except for the Kutupalong-2018 dataset, it has relatively moderate structural similarity (see Fig. 9) and good transfer performance (Fig. 6). Here it is also noted that for this pattern, transfers from other datasets to Kutupalong-2017 do not result in the same conclusion with the exception of Nguenygiel-2018 to Kutupalong-2017 transfers. Although one might expect images taken from the study site at

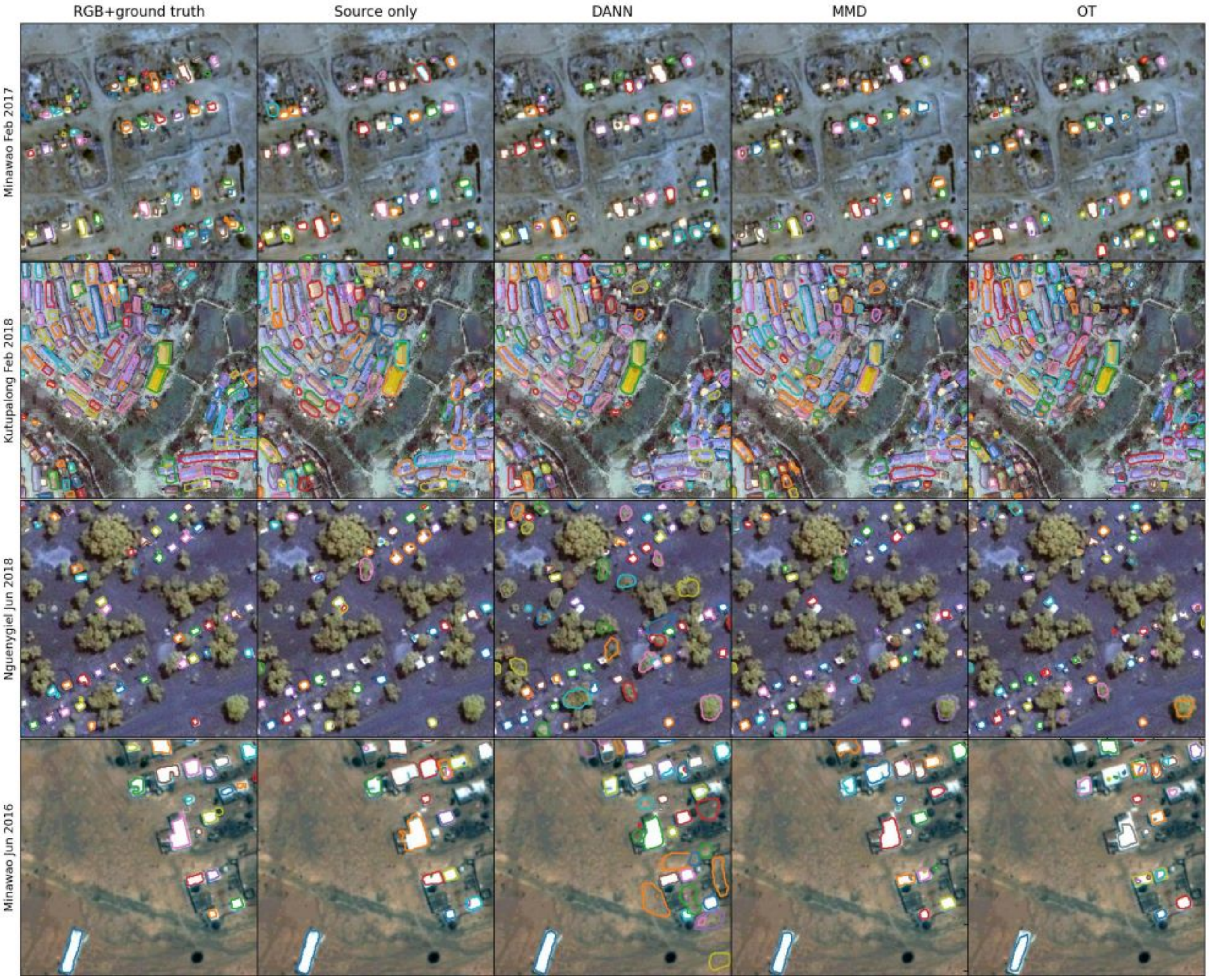

Fig. 6. Spatial plots of inference made using Kutupalong-2017 as source datasets. Each row represents a different target dataset.

different time scales to show a higher image-level similarity, computed SSIM values indicate the opposite. This can be seen from SSIM between Kutupalong-2017 and Kutupalong-2018 images and between Minawao-2016 and Minawao-2017 images, which all are a clear indication of the temporal dynamics of FDP settlements even with a time lag of less than a year. The minimum SSIM is observed between images taken from Kutupalong-2017 and Kutupalong-2018. As can be seen from Fig. 6, a lower transfer performance is also observed in transfers made from Kutupalong-2018 to other datasets. This pattern is very common for MMD and DANN-based transfer strategies.

Second, we look into object-level similarity. Fig. 10 presents object-level domain similarity using selected metrics (shape index, distance to nearest neighbour and dwelling density) for each study site. From object-level domain similarity analysis presented in Fig. 10, almost all datasets have similar shape complexity which is inferred from the shape index. The largest variations are observed for distance to nearest neighbour and dwelling density metrics. The smaller distance to the nearest dwelling could be an indication of a congested spatial place occupancy pattern. Coupled with larger dwelling density, it is straightforward to infer the dwellings exhibit a complex pattern. The distances to the nearest neighbour for the Kutupalong-2017 and 2018 datasets are both lower and almost the same except for the number of dwellings per hectare. Though Kutupalong-2017 and Kutupalong-2018 datasets have relatively similar object-level metrics, if other things remain constant, in all strategies one could expect better domain adaptation transfer performance between these datasets than others. In instance segmentation tasks, if the dwelling objects are closely spaced or packed together it poses an inherent challenge to detect individual dwelling instances which is commonly prevalent in Kutupalong datasets. In densely populated complex dwelling structures, not only the transfer but also instance segmentation with supervised learning is found difficult (a notable example is the Kutupalong-2018 dataset (except transfers with OT approach) as can be seen from Fig. 6 with a low supervised performance of about mAP 0.38. This could be attributed to the difficulty of retrieving individual

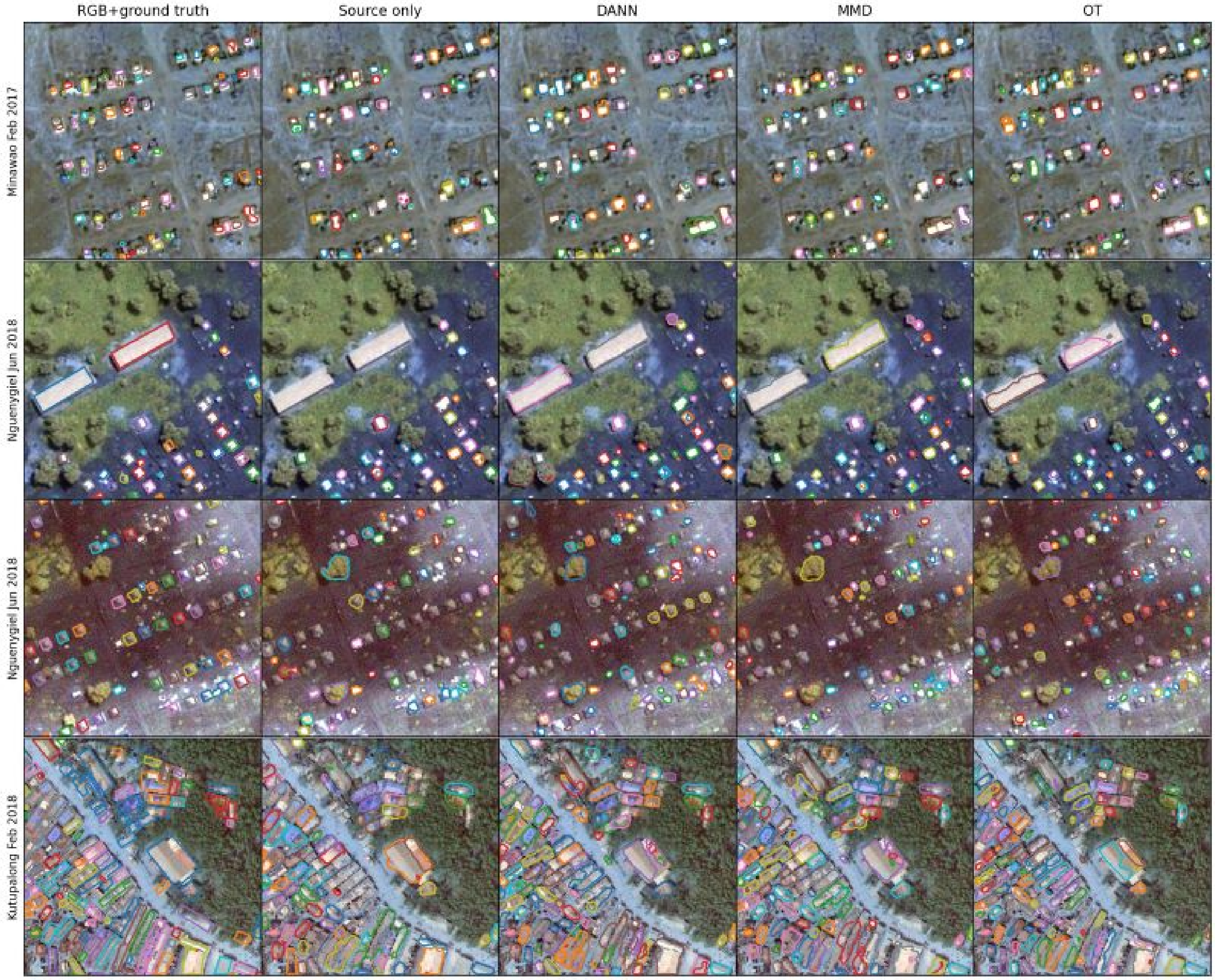

Fig. 7. Performance of adopted approaches in a complex patch. Examples of transfers made from Kutupalong-2017 to Kutupalong-2018, Nguenygiel-2018, Minawao-2016 and Minawao-2017 datasets. Except for the Minawawo-2017 dataset, the source-only transfers have visible false negatives

dwellings from crowded, spectrally diverse and low contrast with background environment (see sample patches in Fig. 1 for Kutupalong dataset).

Finally, we analyse the link between deep feature space similarity and transfer performance. Fig. 8 displays the deep feature spaces obtained from t-SNE. This decision boundary indicates the inliers displayed with filled polygons for each dataset to indicate the deep feature space distances. Table 9 presents GIoU as deep feature space level domain similarity. While there is an association between the transfer performance and image- and object-level similarity metrics of some source-target combinations, transfer performance and feature space-level domain similarity show a rather random pattern. Based on the results presented in Table III, one could expect a higher transfer performance would be achieved between Kutupalong-2017 and Kutupalong-2018 and Minawao-2016 and Minawao-2017 datasets that have minimal GIoU values. Despite their relatively better deep feature space similarity than other combinations, their transfer performance falls short of expectations except for OT-based transfers from Kutupalong-2017 → Kutupalong-2018. This random pattern between deep feature space similarity and transfer performance is observed for other datasets too (see Fig. 5, Table III, and Fig. 8).

TABLE III
GENERALIZED INTERSECTION OVER UNION (GIOU) VALUES MEASURE DEEP-FEATURE SPACE-LEVEL SIMILARITY BETWEEN SOURCE AND TARGET DOMAINS.

	KT17	KT18	MW17	MW16	NG18
Kutupalong-2018	-0.493				
Minawao-2017	-0.626	-0.646			
Minawao-2016	-0.609	-0.781	-0.487		
Nguenygiel-2018	-0.797	-0.799	-0.644	-0.724	
Nduta-2016	-0.742	-0.749	-0.836	-0.822	-0.767

Note: The theoretical minimum value is -1 for extremely distant geometric objects and +1 for perfectly overlapping objects. In general, negative values are indications of non-overlapping feature spaces.

As presented in Table IV, the computed statistical test has

TABLE IV
SIMILARITY METRICS AND UNSUPERVISED DOMAIN ADAPTATION TRANSFER PERFORMANCE ASSOCIATION TEST

	SSIM				GIoU				SHI				NoH				MDN			
Model	NoDA	DANN	MMD	OT	NoDA	DANN	MMD	OT	NoDA	DANN	MMD	OT	NoDA	DANN	MMD	OT	NoDA	DANN	MMD	OT
r	0.69	0.51	0.35	0.50	-0.30	0.14	0.14	-0.08	-0.42	-0.45	-0.31	-0.46	-0.41	-0.12	-0.07	-0.23	-0.35	0.01	0.09	-0.31
p	0.00**	0.05**	0.19	0.05**	0.28	0.63	0.63	0.77	0.11	0.09*	0.25	0.08*	0.27	0.27	0.28	0.18	0.20	0.96	0.76	0.25

SHI, NoH and MDN stands for shape index, number of dwellings per hectare, and mean distance to nearest neighbour
r, p indicates the Pearson correlation coefficient and the test significance (p-value) with a 90% (*) or 95% (**) confidence interval.

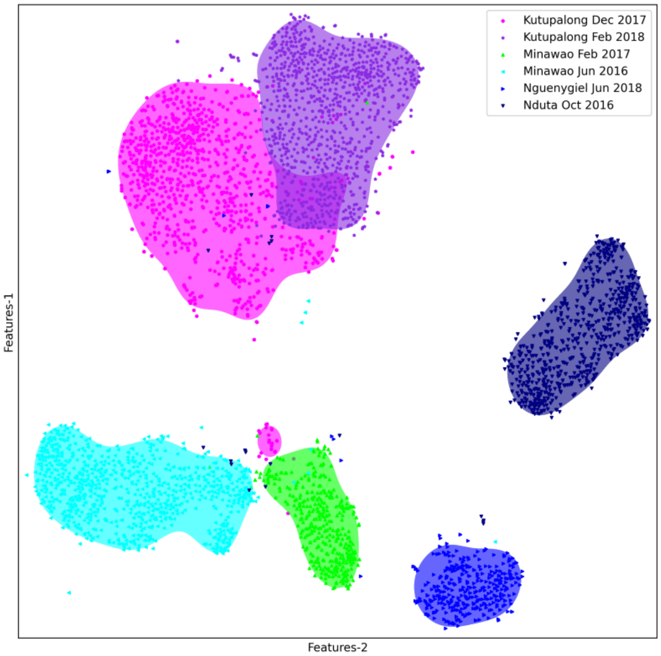

Fig. 8. Visualization of inter and intra-deep feature space disparity as qualitative feature space level domain similarity. Each individual point represents a patch. The closer the points are, the more similar the features are. Filled contours indicate the decision boundaries obtained by one class Support Vector Machine (OC-SVM) between the patch features from one domain and the others.

shown three patterns on the association of transfer performance and domain similarity at different levels. The first pattern is moderate association and statistically significant. This is observed between image level similarity (SSIM) and transfer results from source only (NoDA), DANN and OT strategies. A similar pattern is also observed between one of the object-level similarity metrics (shape index) and the DANN and OT strategies. The second pattern is a relatively moderate to small association but statistically insignificant pattern, which is mostly a pattern for transfers and object level similarity metrics (number of dwellings per hectare, mean distance to nearest neighbour and shape index; except for the OT technique, which is statistically significant). The correlation coefficient values for object-based domain similarity metrics are negative (except for the mean minimum distance to the nearest neighbour with DANN and MMD transfers). This shows when there is a wider gap in object-level domain similarity, it inversely affects the transfer performance. The third pattern is a very low association and at the same time statistically insignificant association. This is a general pattern for deep feature space level domain similarity (GIoU) and all the transfer scenarios.

V. DISCUSSION

Based on the results presented in section IV, we provide in this section a discussion on two main aspects: (1) the performance of the investigated domain adaptation approaches, and (2) the link between domain similarity and transfer performance. In datasets where dwelling characteristics are extremely contrasting in terms of spatial and spectral patterns, domain adaptation helped with the accurate segmentation of dwelling objects. In a few cases (e.g., Kutupalong-2018 as target dataset with the exception of OT strategy), low performance is observed in both the supervised setting and the transfer scenarios. The same pattern is also observed in transfers where Kutupalong-2017 is a target dataset except for transfers Nguenygiel-2018 → Kutupalong-2017.

It should also be noted that the initial pre-training performance could have an impact on the transfer performance. As can be seen from Fig. 5, if the model performance is not good during the initial pre-training (see the red dashed lines), the transfer performance is limited (e.g., this can be seen from transfers where the Kutuplong-2018 dataset served as the source dataset). On the other hand, if the model performed well during pre-training on the source dataset (e.g., Nguenygiel-2018 and Kutupalong-2017 datasets), it achieves good transfer performance on target datasets. As observed in Fig. 5, in a few cases domain adaptation strategies perform lower than transfers done without any domain adaptation (source-only transfers). Initial pre-training on the source dataset could yield a learned representation which could be far from the target dataset representations. Undertaking joint domain adaptation could end up with different source and target distributions. A negative transfer could happen if the source domain has a different data distribution than the target domain. Both situations demand precaution in the trade-off between negative transfers [78], [79] and advantages of pretraining. Even though three of the implemented domain adaptation strategies fall under deep domain alignment unsupervised domain adaptation strategies, MMD and DANN tackle domain gaps only on feature spaces while OT includes label space alignment. In this aspect, their performance has also a clear pattern. DANN and MMD achieved closer results. Though in most cases OT approach performed better than DANN and MMD, in a few cases they achieved better results than OT. Transfers Nguenygiel-2018 → Kutupalong-2017, Kutupalong-2017 → Minawao-2016, Nguenygiel-2018 → Nduta-2016 are notable examples.

From the visual comparison, over-predictions are observed on outputs obtained from DANN-based transfer. Dismantled building basements, tree stands and footpaths are confused

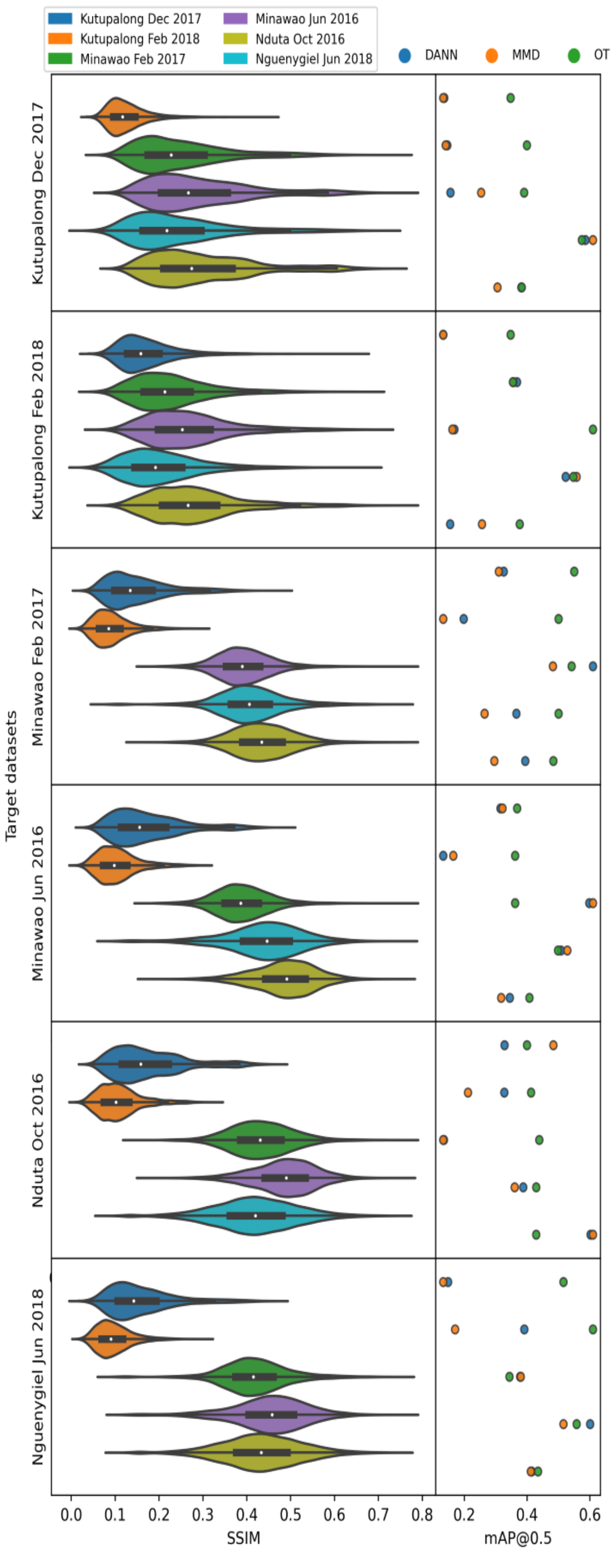

Fig. 9. Structural Similarity Index values across the datasets (first column) as a measure of image-label domain similarity together with transfer performance (second column). The violin plots were scaled to width.

with dwellings. In a nutshell, transfer performance could be associated with the implemented domain transfer approach, pre-training, and nature of the source and target dataset [80].

It should also be noted that the evaluation metrics reported in the study are based on standard COCO evaluation metrics where false positives are determined using an IoU value of 0.5 on predictions obtained with a score threshold of 0.3. As can be seen from Fig. 12, over-predictions could contribute to reported low mAP values, especially on patches dominated by relatively bigger dwellings with spectrally diverse rooftops caused by rusting, the use of different materials, and the changes in the natural morphology of dwellings. For operational use, post-processing of over-predictions by using an objectness score (category score after matrix NMS) could improve the final segmentation product. For multiple predictions of small dwellings predicted with a large objectness score merging them with bigger dwellings could helped to refine the final result (Fig. 12). In addition to this, non-dwelling features like footpaths and small individual trees were segmented as dwelling objects which also increases false positive predictions. In the Nduta datasets, detached individual standing trees were detected as dwellings, thus increasing the false positive rate in the prediction (Fig. 6). Based on the availability of the Near Infrared (NIR) band, the issue could also be mitigated by using a threshold on vegetation indices like Normalized Difference Vegetation Index (NDVI). Post-processing is not easy and comes at the cost of other errors. While using either objectness score or NDVI thresholds, caution should also be paid as it would play with trade-offs of increasing false negatives.

As can be seen from the result section, Kutupalong 2017 and 2018 datasets, which are taken from the same area, have a relatively small feature space discrepancy (Fig. 8 and Table III) and relatively similar object-level domain similarities (Fig. 10). Hence, in all strategies, one could expect better transfer performance between those datasets than others though the results show the opposite. This clue could be an indication that image-level similarity has a higher influence on transfer performance than object and deep feature level similarity. Except for domain transfers made with the MMD approach, we observe that image-level domain similarity has a statistically significant moderate association with domain transfer performance. Even though there is an extreme influx of settlers which caused camp expansion and dwelling densification [1], changes in object level similarity of dwellings are negligible. This is mainly because new dwellings were built in empty patches at the outer expanses of the camp without changing the dwelling size (Table II) and the space occupancy pattern. Changes are observed in background environment, rooftop colours because of changes in the rooftop material, and natural morphology like rusting and dust (see the first two rows of Fig. 1 and Fig. 11). This causes strong image-level domain discrepancy and respective low domain adaptation performance. Overall, a better association could easily be established with image-level similarity than feature space and selected object-level similarity (shape index) (see Table IV).

Given that image and dwelling object-level domain similarities have shown meaningful clues on domain transfer per-

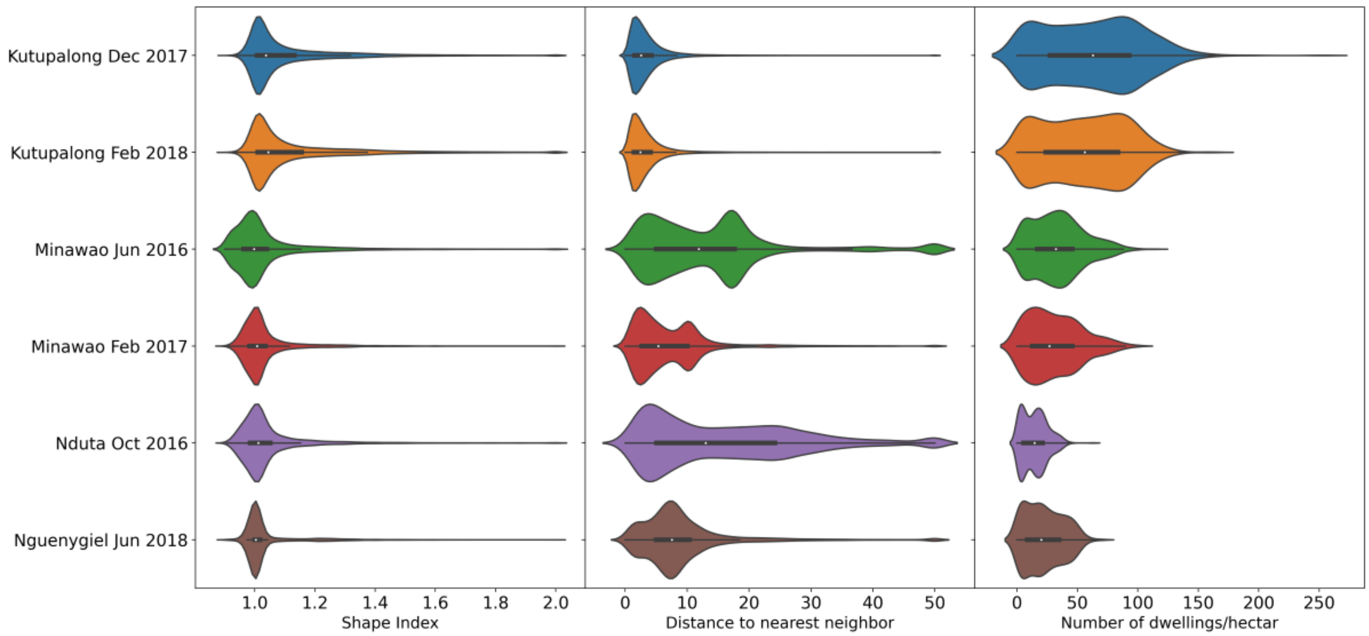

Fig. 10. Selected dwelling object characteristics as a measure of object label domain similarity. Violin plots were scaled to width. Outlier values from shape index and distance to the nearest-neighbour metrics are excluded for visual quality.

formance, during emergency response, existing source datasets could be filtered out using the corresponding similarity metrics to undertake unsupervised domain adaptation with unlabeled target dataset. Though not in emergency response operations, the use of domain similarity information for accurate domain adaptation is reported in a handful of studies. For example, the relevance of source dataset selection is explained in [81] using machine-learning-based domain adaptation for regression tasks given many remotely sensed source datasets and a single target dataset. The authors have implemented histogram matching of spectral features and reported Pearson correlation and Hellinger distance as effective metrics for domain similarity. In non-Earth Observation datasets, [82] has implemented distance-based metrics in deep feature space for source selection while [83] has applied domain reweighting based on inter-domain similarity for multi-source domain adaptation. Their approach could be adapted to any kind of domain similarity metrics, e.g., in our case with image-level similarity. There is a potential to extend their study for dwelling extraction. Please note that compared to these studies that are based only on spectral features, we consider different domain-similarity metrics, which complement the former works.

VI. Conclusions

To bypass data-intensive training, transferring, and fine-tuning, three domain adaptation approaches, namely domain adversarial, domain discrepancy, and domain mapping, were tested for unsupervised instance segmentation of dwelling objects from very high-resolution satellite imagery using the single-stage instance segmentation model SOLO. Generally speaking, in most source and target combinations, unsupervised domain adaptation brought a performance improvement with a large margin compared to transferring the model without any domain adaptation. Though this is the general scenario, the under-performance of domain adaptation approaches has also occurred, which is dataset-dependent. While comparing investigated unsupervised domain adaptation approaches, domain-adversarial (DANN) and -discrepancy (MMD) adaptation techniques have consistently shown competitive transfer performances. In most cases, optimal transport has shown better transfer performance than MMD and DANN strategies. This better performance could be attributed to the use of label space alignment, other than MMD and DANN which only looks at domain similarity in the feature spaces. It should also be noted that this is the first time that OT was used for instance segmentation. We have followed a more simplified label space alignment term that, the label space alignment loss needs redesigning for even better and robust performance. From a thorough visual analysis of computed domain similarity and statistical test for association between domain similarity and transfer performance, image level domain similarity (SSIM) has shown better associated pattern with unsupervised domain adaptation performance. Though object-level similarity has shown from moderate to low association, except for the shape index on OT-based transfer, the association is not statistically significant. For the instance segmentation transferring from datasets with sparsely populated, regular patterns and small-sized buildings, towards dense and mixed spectral characteristics did not provide good results. It is also found hard to link the domain adaptation performance with the deep feature space as it shows statistically insignificant low association statistic and random pattern. As object-level domain similarity also needs supervision, and thus labels in both domains, image-level domain similarity could be an ideal domain similarity metric to select labeled source datasets for a transfer in an

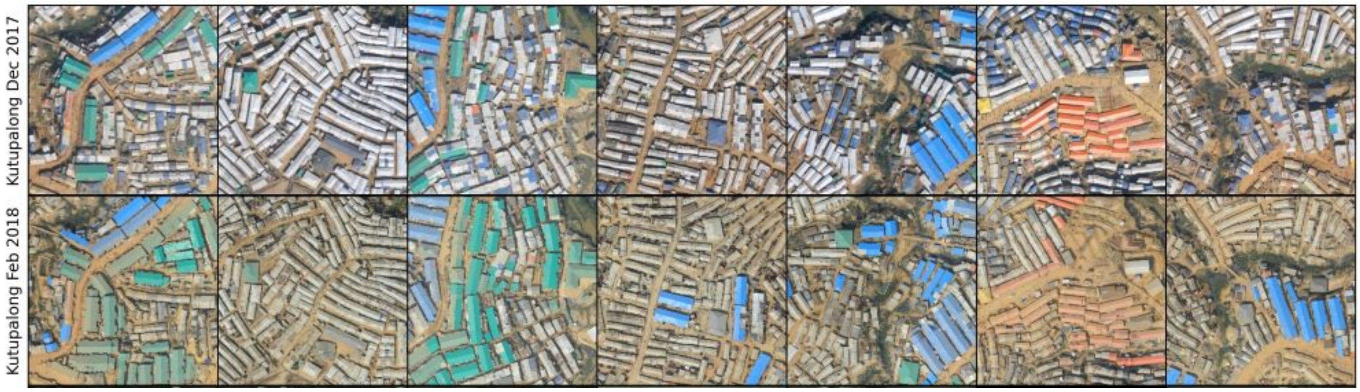

Fig. 11. Example of changes in spectral characteristics (color) through time for Kutupalong dataset.

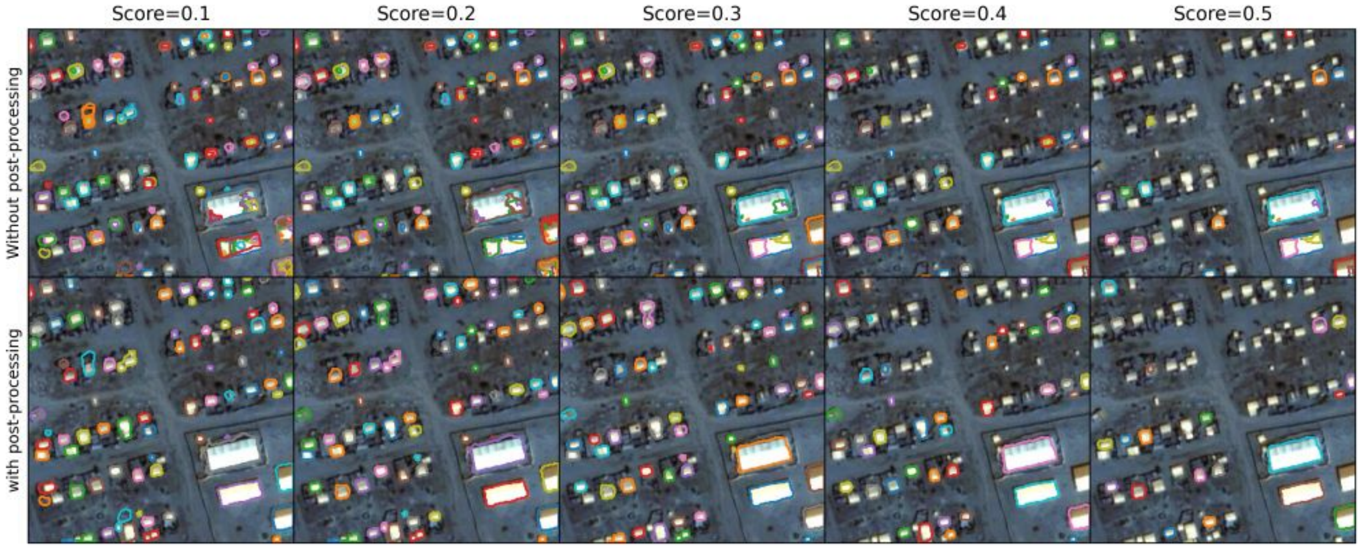

Fig. 12. Output sensitivity for modified score threshold (objectness score) and role of post-processing. Results are taken from transfers Kutupalong-2017 → Minawao-2017 using the Maximum Mean Discrepancy domain approach.

unlabelled target dataset during humanitarian emergency response. Hence, it is reasonable to look into domain alignment losses which could account for features from early layers of the model and use of image-level domain similarity as auxiliary information for unsupervised loss and batch wise loss reweighting during multi-source domain adaptation. As the current study has focused on binary instance segmentation of dwelling from paired source-target datasets, it will be extended to multi-source and multi-class instance segmentation tasks using either domain reweighting or proper selection of a few source datasets based on domain similarity.

Acknowledgements

Authors from PLUS are funded by Christian Doppler Research Association, Doctors without Borders-Section Austria, the Austrian Federal Ministry of Labour and Economy and the National Foundation for Research, Technology and Development. We are also grateful for the partial financial support obtained from the Erasmus+ mobility program for the first author's stays in the University of South Brittany (UBS), Environment Observation with Complex Imagery (OBELIX) research lab. Finally, we thank Javiera Castillo-Navarro for her inspiring inputs on deep feature space visualization. We are also very grateful to anonymous reviewers for their insightful comments that enabled us to improve the manuscript.

Appendix

In this section, we present test performance from different instance segmentation models (one-stage and two-stage) combined with different feature extractor networks. The main intention of this analysis was to select an optimal model with good speed and low memory requirements without compromising segmentation performance. Table V shows that SOLO-v2 with a ResNet-50 feature extractor meets this objective: it has high accuracy in a supervised setting without compromising speed and memory usage.

TABLE V
TEST PERFORMANCE FOR THE MINAWAO-2016 DATASET OBTAINED FOR DIFFERENT INSTANCE SEGMENTATION MODELS ASSOCIATED WITH DIFFERENT FEATURE EXTRACTORS.

Model	Feature extractor	Params	Slow (steps/s)	Fast (steps/s)	Mean ± std	Test time(tasks/s)	mAP@.5± std
SOLOv2	Resnet50	46.01×10^6	0.3520	0.3289	0.3400 ± 0.0041	8	**0.8233** ± 0.0117
SOLO-v2	ResNet101	65×10^6	0.3873	0.3615	0.3736 ± 0.0064	10	0.7881 ± 0.0214
SOLO-v2	ResNet101_dcn	68.14×10^6	0.7273	0.6844	0.7116 ± 0.0083	12	0.7682 ± 0.0261
SOLO-v2	SwinT	49.63×10^6	0.6137	0.5981	0.6048 ± 0.0032	12	0.7949 ± 0.0373
SOLO-v2	SwinS	70.95×10^6	0.6120	0.5920	0.6036 ± 0.0049	13	0.7884 ± 0.0035
SOLO-v2	SwinB	108.98×10^6	0.7293	0.7181	0.7240 ± 0.724	17	0.8139 ± 0.0041
SOLO-v2	SwinL	217.48×10^6	0.9924	0.9805	0.9870 ± 0.0027	22	**0.8099** ± 0.0532
SOLO-v1	ResNet50	39.62×10^6	0.4367	0.4239	0.4301 ± 0.0036	9	0.7729 ± 0.0053
SOLO-v1	ResNet101	58.62×10^6	0.5466	0.5210	0.5290 ± 0.0040	11	0.7919 ± 0.0082
Mask RCNN	ResNet50	43.75×10^6	0.3222	0.3060	0.3159 ± 0.0032	12	0.7872 ± 0.0012
MasK RCNN	ResNet101	62.74×10^6	0.4038	0.3877	0.3974 ± 0.0031	14	0.7872 ± 0.0012
Mask RCNN	ResNet50 with PointRend	55.54×10^6	0.2944	0.2793	0.2866 ± 0.0029	12	0.8021 ± 0.0015
Mask RCNN	ResNet101 with PointRend	74.48×10^6	0.3462	0.2853	0.2936 ± 0.0085	15	0.8132 ± 0.0051
Mask RCNN	SwinT	47.37×10^6	0.4099	0.3867	0.4007 ± 0.004	26	0.7952 ± 0.0051
Mask RCNN	SwinS	68.69×10^6	0.4583	0.4330	0.4465 ± 0.0058	27	0.8052 ± 0.0012
Mask RCNN	SwinB	106.72×10^6	0.4931	0.4520	0.4654 ± 0.0083	36	0.8101 ± 0.001
Mask RCNN	SwinL	215.22×10^6	0.7321	0.5520	0.6121 ± 0.0031	39	0.8080 ± 0.0031
YOLACT	ResNet50	34.73×10^6	0.1247	0.1189	0.1217 ± 0.0012	19	0.7759 ± 0.0052
YOLACT	ResNet101	53.73×10^6	0.1798	0.1768	0.1781 ± 0.0009	20	0.7766 ± 0.0047

REFERENCES

[1] S. Benz, H. Park, J. Li, D. Crawl, J. Block, M. Nguyen, and I. Altintas, "Understanding a rapidly expanding refugee camp using convolutional neural networks and satellite imagery," in *2019 15th International Conference on eScience (eScience)*. IEEE, 2019, pp. 243–251.

[2] A. Braun, F. Fakhri, and V. Hochschild, "Refugee camp monitoring and environmental change assessment of kutupalong, bangladesh, based on radar imagery of sentinel-1 and alos-2," *Remote Sensing*, vol. 11, no. 17, p. 2047, 2019.

[3] B. Tomaszewski, S. Tibbets, Y. Hamad, and N. Al-Najdawi, "Infrastructure evolution analysis via remote sensing in an urban refugee camp–evidence from za'atari," *Procedia engineering*, vol. 159, pp. 118–123, 2016.

[4] A. Hadzic, G. Christie, J. Freeman, A. Dismer, S. Bullard, A. Greiner, N. Jacobs, and R. Mukherjee, "Estimating displaced populations from overhead," in *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2020, pp. 1121–1124.

[5] N. Ahmed, N. A. Diptu, M. S. K. Shadhin, M. A. F. Jaki, M. F. Hasan, M. N. Islam, and R. M. Rahman, "Artificial neural network and machine learning based methods for population estimation of rohingya refugees: Comparing data-driven and satellite image-driven approaches," *Vietnam Journal of Computer Science*, vol. 6, no. 04, pp. 439–455, 2019.

[6] M. Hagenlocher, S. Lang, and D. Tiede, "Integrated assessment of the environmental impact of an idp camp in sudan based on very high resolution multi-temporal satellite imagery," *Remote Sensing of Environment*, vol. 126, pp. 27–38, 2012.

[7] S. Wang, E. So, and P. Smith, "Detecting tents to estimate the displaced populations for post-disaster relief using high resolution satellite imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 36, pp. 87–93, 2015.

[8] K. Spröhnle, D. Tiede, E. Schoepfer, P. Füreder, A. Svanberg, and T. Rost, "Earth observation-based dwelling detection approaches in a highly complex refugee camp environment—a comparative study," *Remote Sensing*, vol. 6, no. 10, pp. 9277–9297, 2014.

[9] K. Spröhnle, E.-M. Fuchs, and P. A. Pelizari, "Object-based analysis and fusion of optical and sar satellite data for dwelling detection in refugee camps," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 5, pp. 1780–1791, 2017.

[10] D. Tiede, S. Lang, D. Hölbling, and P. Füreder, "Transferability of obia rulesets for idp camp analysis in darfur," *Proceedings of the GEOBIA*, pp. 1–6, 2010.

[11] S. Ergünay, F. Kahraman, and H. F. Ates, "Automated detection of refugee/idp tents from satellite imagery using two-level graph cut segmentation," *Imaging and Mapping for Disaster Management*, 2013.

[12] W. Gu, S. Bai, and L. Kong, "A review on 2d instance segmentation based on deep neural networks," *Image and Vision Computing*, p. 104401, 2022.

[13] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 12, pp. 4338–4364, 2020.

[14] W. Sirko, S. Kashubin, M. Ritter, A. Annkah, Y. S. E. Bouchareb, Y. Dauphin, D. Keysers, M. Neumann, M. Cisse, and J. Quinn, "Continental-scale building detection from high resolution satellite imagery," *arXiv preprint arXiv:2107.12283*, 2021.

[15] T. G. Tiecke, X. Liu, A. Zhang, A. Gros, N. Li, G. Yetman, T. Kilic, S. Murray, B. Blankespoor, E. B. Prydz *et al.*, "Mapping the world population one building at a time," *arXiv preprint arXiv:1712.05839*, 2017.

[16] J. Van Den Hoek and H. K. Friedrich, "Satellite-based human settlement datasets inadequately detect refugee settlements: A critical assessment at thirty refugee settlements in uganda," *Remote Sensing*, vol. 13, no. 18, p. 3574, 2021.

[17] O. Ghorbanzadeh, D. Tiede, L. Wendt, M. Sudmanns, and S. Lang, "Transferable instance segmentation of dwellings in a refugee camp-integrating cnn and obia," *European Journal of Remote Sensing*, vol. 54, no. sup1, pp. 127–140, 2021.

[18] Y. Lu, K. Koperski, C. Kwan, and J. Li, "Deep learning for effective refugee tent extraction near syria–jordan border," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 8, pp. 1342–1346, 2020.

[19] J. A. Quinn, M. M. Nyhan, C. Navarro, D. Coluccia, L. Bromley, and M. Luengo-Oroz, "Humanitarian applications of machine learning with remote-sensing data: Review and case study in refugee settlement mapping," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2128, p. 20170363, 2018.

[20] O. Ghorbanzadeh, D. Tiede, Z. Dabiri, M. Sudmanns, and S. Lang, "Dwelling extraction in refugee camps using cnn-first experiences and lessons learnt." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 42, no. 1, 2018.

[21] G. W. Gella, L. Wendt, S. Lang, D. Tiede, B. Hofer, Y. Gao, and A. Braun, "Mapping of dwellings in idp/refugee settlements from very high-resolution satellite imagery using a mask region-based convolutional neural network," *Remote Sensing*, vol. 14, no. 3, p. 689, 2022.

[22] D. Tiede, G. Schwendemann, A. Alobaidi, L. Wendt, and S. Lang, "Mask r-cnn-based building extraction from vhr satellite data in operational humanitarian action: An example related to covid-19 response in khartoum, sudan," *Transactions in GIS*, vol. 25, no. 3, pp. 1213–1227, 2021.

[23] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE geoscience and remote sensing magazine*, vol. 4, no. 2, pp. 41–57, 2016.

[24] M. Zhang, H. Singh, L. Chok, and R. Chunara, "Segmenting across places: The need for fair transfer learning with satellite imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2916–2925.

[25] M. Xu, M. Wu, K. Chen, C. Zhang, and J. Guo, "The eyes of the gods: A survey of unsupervised domain adaptation methods based on remote sensing data," *Remote Sensing*, vol. 14, no. 17, p. 4380, 2022.

[26] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 105–109, 2015.

[27] R. Naushad, T. Kaur, and E. Ghaderpour, "Deep transfer learning for land use and land cover classification: A comparative study," *Sensors*, vol. 21, no. 23, p. 8083, 2021.

[28] W. Yang, X. Zhang, and P. Luo, "Transferability of convolutional neural network models for identifying damaged buildings due to earthquake," *Remote Sensing*, vol. 13, no. 3, p. 504, 2021.

[29] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," in *International Conference on Learning Representations*, 2022.

[30] X. Liu, C. Yoo, F. Xing, H. Oh, G. El Fakhri, J.-W. Kang, J. Woo *et al.*, "Deep unsupervised domain adaptation: A review of recent advances and perspectives," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[31] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[32] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[33] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[34] Y. Ning, J. Peng, L. Sun, Y. Huang, W. Sun, and Q. Du, "Adaptive local discriminant analysis and distribution matching for domain adaptation in hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4797–4808, 2022.

[35] C. Yu, C. Liu, H. Yu, M. Song, and C.-I. Chang, "Unsupervised domain adaptation with dense-based compaction for hyperspectral imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 12 287–12 299, 2021.

[36] C. Zhang, Y. Feng, L. Hu, D. Tapete, L. Pan, Z. Liang, F. Cigna, and P. Yue, "A domain adaptation neural network for change detection with heterogeneous optical and sar remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 109, p. 102769, 2022.

[37] J. Liu, W. Xuan, Y. Gan, Y. Zhan, J. Liu, and B. Du, "An end-to-end supervised domain adaptation framework for cross-domain change detection," *Pattern Recognition*, vol. 132, p. 108960, 2022.

[38] J. Guo, J. Yang, H. Yue, and K. Li, "Unsupervised domain adaptation for cloud detection based on grouped features alignment and entropy minimization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[39] H. Chen, H. Zhang, G. Yang, S. Li, and L. Zhang, "A mutual information domain adaptation network for remotely sensed semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[40] N. Bengana and J. Heikkilä, "Improving land cover segmentation across satellites using domain adaptation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1399–1410, 2020.

[41] E. Othman, Y. Bazi, F. Melgani, H. Alhichri, N. Alajlan, and M. Zuair, "Domain adaptation network for cross-scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4441–4456, 2017.

[42] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sensing*, vol. 10, no. 1, p. 144, 2018.

[43] Y. Cai, Y. Yang, Q. Zheng, Z. Shen, Y. Shang, J. Yin, and Z. Shi, "Bifdanet: Unsupervised bidirectional domain adaptation for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 14, no. 1, p. 190, 2022.

[44] J. Peng, Y. Huang, W. Sun, N. Chen, Y. Ning, and Q. Du, "Domain adaptation in remote sensing image classification: A survey," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9842–9859, 2022.

[45] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia *et al.*, "A review of single-source deep unsupervised visual domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 473–493, 2020.

[46] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Advances in Neural information processing systems*, vol. 33, pp. 17 721–17 732, 2020.

[47] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Solo: A simple framework for instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8587–8601, 2021.

[48] K. Vogt, A. Paul, J. Ostermann, F. Rottensteiner, and C. Heipke, "Unsupervised source selection for domain adaptation," *Photogrammetric Engineering & Remote Sensing*, vol. 84, no. 5, pp. 249–261, 2018.

[49] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[50] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty, "Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 447–463.

[51] A. Ackaouy, N. Courty, E. Vallée, O. Commowick, C. Barillot, and F. Galassi, "Unsupervised domain adaptation with optimal transport in multi-site segmentation of multiple sclerosis lesions from mri data," *Frontiers in computational neuroscience*, vol. 14, p. 19, 2020.

[52] R. Flamary, N. Courty, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 1, 2016.

[53] X. Huang, L. Ren, C. Liu, Y. Wang, H. Yu, M. Schmitt, R. Hänsch, X. Sun, H. Huang, and H. Mayer, "Urban building classification (ubc)-a dataset for individual building detection and classification from satellite imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1413–1421.

[54] X. Liang, L. Lin, Y. Wei, X. Shen, J. Yang, and S. Yan, "Proposal-free network for instance-level object segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2978–2991, 2017.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[56] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[57] R. Liu, J. Lehman, P. Molino, F. Petroski Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," *Advances in neural information processing systems*, vol. 31, 2018.

[58] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[59] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, and Y. Pan, "Rethinking dice loss for medical image segmentation," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 851–860.

[60] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[61] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[62] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier *et al.*, "Pot: Python optimal transport," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 3571–3578, 2021.

[63] C. M. Gevaert and M. Belgiu, "Assessing the generalization capability of deep learning networks for aerial image classification using landscape metrics," *International Journal of Applied Earth Observation and Geoinformation*, vol. 114, p. 103054, 2022.

[64] M. Bosch, "Pylandstats: An open-source pythonic library to compute landscape metrics," *PloS one*, vol. 14, no. 12, p. e0225734, 2019.

[65] 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," *Machine Learning*, vol. 111, pp. 3125–3160, 2022.

[66] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[67] R. Gove, L. Cadalzo, N. Leiby, J. M. Singer, and A. Zaitzeff, "New guidance for using t-sne: Alternative defaults, hyperparameter selection automation, and comparative evaluation," *Visual Informatics*, vol. 6,

no. 2, pp. 87–97, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468502X22000201

[68] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[69] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[70] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

[71] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[72] N. S. Detlefsen, J. Borovec, J. Schock, A. H. Jha, T. Koker, L. Di Liello, D. Stancl, C. Quan, M. Grechkin, and W. Falcon, "Torchmetrics-measuring reproducibility in pytorch," *Journal of Open Source Software*, vol. 7, no. 70, p. 4101, 2022.

[73] S. Lang, L. Wendt, D. Tiede, Y. Gao, V. Streifender, H. Zafar, A. Adebayo, G. Schwendemann, and P. Jeremias, "Multi-feature sample database for enhancing deep learning tasks in operational humanitarian applications," in *GI_Forum*, 2021, vol. 9, no. 1, pp. 209–219.

[74] DigitalGlobe, "Core imagery products guide," in *DigitalGlobe*, 2021. [Online]. Available: https://www.geosoluciones.cl/documentos/worldview/DigitalGlobe-Core-Imagery-Products-Guide.pdf

[75] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," Jan. 4 2000, uS Patent 6,011,875.

[76] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[77] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, pp. 98–136, 2015.

[78] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[79] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, "Characterizing and avoiding negative transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 293–11 302.

[80] D. Kim, K. Wang, S. Sclaroff, and K. Saenko, "A broad study of pre-training for domain generalization and adaptation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*. Springer, 2022, pp. 621–638.

[81] C. Geiß, A. Rabuske, P. A. Pelizari, S. Bauer, and H. Taubenböck, "Selection of unlabeled source domains for domain adaptation in remote sensing," *Array*, vol. 15, p. 100233, 2022.

[82] L. R. Schultz, M. Loog, and P. M. Esfahani, "Distance based source domain selection for sentiment classification," *arXiv preprint arXiv:1808.09271*, 2018.

[83] K. Bascol, R. Emonet, and E. Fromont, "Improving domain adaptation by source selection," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3043–3047.

Getachew Workineh Gella is at the University of Salzburg, Department of Geoinformatics working toward his PhD. He has a broader interest in the integration of earth observation and deep learning for automatic information retrieval and spatio-temporal transferability of deep learning models.

Charlotte Pelletier is an Associate Professor in computer science at Univ. Bretagne Sud, Vannes, France. She conducts her research at the Institute for Research in Information Technology and Random Systems (IRISA), Vannes, France. Her research interests include time series analysis, tree-based approaches, and deep learning techniques with applications to Earth observations. She chairs an ISPRS working group on geospatial temporal data understanding (2022-2024) and co-chairs the IAPR Technical Committee 7 on remote sensing and mapping (2021-2024). She is the coordinator of the GeoData Science track within the Erasmus Mundus Joint Master Degree named Copernicus Master in Digital Earth.

Sébastien Lefèvre is a Professor of Computer Science at the Univ. Bretagne Sud, Vannes, France. He conduct his research in the OBELIX team he founded in 2013 (and let until 2021) at the Institute for Research in IT and Random Systems (IRISA). He is coordinating the GeoData Science track within the Erasmus Mundus Joint Master Degree "Copernicus Master in Digital Earth". His main research topics are in efficient remote sensing data analysis and deep learning for Earth observation.

Lorenz Wendt is a remote sensing and GIS specialist and has been part of the humanitarian remote sensing team at the Department of Geoinformatics, University of Salzburg, since 2013. He has been involved in a variety of EO/GI projects in a humanitarian and developments contexts, including water exploration, water infrastructure planning, population estimation, 3D data analysis, and data preparedness, as researcher or consultant.

Dirk Tiede is Associate Professor and Deputy Head of the Department of Geoinformatics, University of Salzburg, Austria, and co-head of the research lab EO Analytics. His research focuses on methodological developments in image analysis using optical EO data, object-based methodologies and process automation in the context of Big EO data analysis. Research fields include environmental monitoring and support of humanitarian relief operations, for which he received the Christian-Doppler-Award of the Federal State of Salzburg in 2014.

Stefan Lang is an Associate Professor at the University of Salzburg specialist in GIS and Remote Sensing is Vice Dean of the Faculty for Digital and Analytical Sciences, and leads the Earth Observation division at the Department of Geoinformatics. He is head of the Christian-Doppler Laboratory(GEOHUM) with research interests in GeoAI, OBIA, multi-source data integration and assimilation, spatial analysis, and regionalization. He initiated a range of collaborative research projects with industry and NGOs and has coordinated a Horizon 2020 project in support of the Copernicus Academy. He chairs the Erasmus+ Joint Master Copernicus in Digital Earth and academic coordinator of the Erasmus+ Sector Skills Alliances project EO4GEO.