Align and Segment: Unsupervised Learning for Building Segmentation From Misaligned Labels

Venkanna Babu Guthula[1], Oswin Krause[1], Dimitri Gominski[1], Hui Zhang[1], Johan Mottelson[2], Ankit Kariryaa[1], Nico Lang[1], and Christian Igel[1]

[1] University of Copenhagen, Copenhagen, Denmark
[2] Royal Danish Academy, Copenhagen, Denmark
{vegu, igel}@di.ku.dk

Abstract. Supervised learning for image segmentation typically requires spatially aligned image and label sets. When images and labels originate from different sources, the pairing may be misaligned, which can significantly deteriorate the performance of the learned models. This is especially common in remote sensing, when aerial or satellite images are co-registered with labels from another source (e.g., OpenStreetMap). In this work, we propose a novel approach for training on misaligned labels, where we simultaneously learn the label alignment. Our align and segment (AnS) approach builds on the spatial transformer module to transform the misaligned labels using an affine transformation to provide a better learning target for a canonical semantic segmentation network. We prevent shortcut learning of misaligned labels in these semantic segmentation networks through a self-supervised regularization loss and show that it is complementary to data augmentation, especially for systematically misaligned training data. A decisive characteristic of our AnS approach is that it learns *without requiring any "golden" labels*. We experimentally show on both synthetic and real-world data from different cities that our approach enables high-quality building segmentation and precise label-image alignment at the same time. Code and derived datasets are available at `https://github.com/venkanna37/align-and-segment`.

Keywords: Segmentation · Label noise · Registration · Unsupervised learning · Remote sensing

1 Introduction

Label noise is a ubiquitous problem in computer vision. In semantic segmentation such noise can either be caused by imprecise annotations or by misalignments when data from different sources are combined. When training a segmentation network on such noisy labels, it is prone to predict the noisy or misaligned labels given enough capacity. The network's performance can sharply deteriorate if the noise is too high [20]. This issue is even worse when the model is trained on systematic label noise, for example, on biased misalignments.

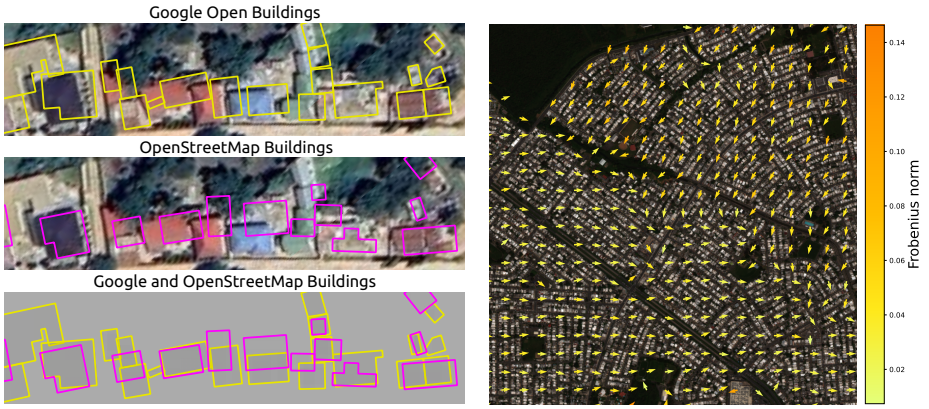

Fig. 1: Misalignment of building footprints with independent imagery. Data e.g. from Google Open Buildings [24] or OpenStreetMap (OSM) [21] neither align with independent satellite imagery nor with each other. The right figure shows our estimated misalignment of OSM data, highlighting the locally systematic label noise. The arrows indicate estimated translations, with color representing the displacement magnitude.

In remote sensing, where images (e.g., satellite or drone imagery) and labels (e.g., crowd-sourced or field-collected) often originate from different sources, label noise is particularly common, especially spatial misalignment of images and label masks. For example, overlaying OpenStreetMap (OSM) [21] building footprints against other public building maps reveals *large and systematic* local transformations of several pixels (Fig. 1). This bottleneck has led to an underutilization of large swaths of hand-curated data such as OSM for training deep learning models at scale and in regions where high-quality building data are scarce.

At first glance, jointly modeling label alignment and segmentation may appear to be a straightforward solution. However, such a disentanglement is nontrivial and a naive implementation in two sub-networks is prone to fail, due to the ability of segmentation networks to simply learn the misaligned labels. An alternative popular approach is to view the problem as a multimodal image alignment task and to estimate the transformation between the image and the label maps using feature maps learned for the different modalities [8, 9, 28]. However, these approaches usually assume that the observed misalignments in the dataset are small and unbiased. In remote sensing these assumptions often do not hold, as can be seen in Fig. 1, where large and biased misalignments of images can be introduced by erroneous orthorectification (i.e., projecting images onto terrain) and other satellite image preprocessing steps as well as errors made in the land registry offices.

In this work, we propose a learning-based approach for dealing with misaligned labels in segmentation without requiring any "golden" (clean, ground-truth) labels. Our solution involves several components of regularization, in-

cluding a new loss that allows two sub-networks to learn label alignment and segmentation respectively. We focus on the segmentation of building footprints from high-resolution satellite imagery, an essential resource for applications such as urban planning as well as disease and disaster risk assessment in informal settlements [1]. Especially for the latter, curated data is rare, as benchmark datasets overwhelmingly focus on urban areas in Europe and the USA [2,19]. While open label maps, such as OSM polygons, exist for these regions, they require extensive alignment and correction by human experts for reliable training [11], which limits the scalability of this valuable label source.

The main contribution of this work is a method for unsupervised learning of multi-modal alignment between image modalities and segmentation maps, without the need for *any* golden labels and without assuming unbiased transformations between the modalities. We summarize our contributions as follows:

1. We propose *Align and Segment (AnS)* as a methodology that can be combined with any existing semantic segmentation network to directly learn from misaligned labels *without using any golden labels*. For this, we propose a combination of a self-consistency loss and data augmentations to directly mitigate the learning of misaligned segmentation maps and biased transformations.
2. We empirically show that AnS applied to building footprint alignment generalizes across different cities and can learn alignment and segmentation for a large range of noise levels and diverse label characteristics, such as building sizes and densities.
3. We compared our method with multiple baselines of supervised and unsupervised methods, and evaluated on real-world and synthetic datasets.

2 Related work

Label noise is a common issue in computer vision, and has become topical in the context of deep learning where models are trained on databases of increasing scale but decreasing curation efforts. For image classification, popular research directions include noise-robust loss functions [30], tailored regularization [16] and label correction [13]. Extending these ideas to semantic segmentation is straight-forward [29,31], but under the unrealistic assumption that label noise is independent and identically distributed among pixels [26]. Accordingly, recent work has tackled label noise in semantic segmentation specifically, with noise modeling [26] or by exploiting the phenomenon of noise robustness in early learning stages [15].

Misaligned supervision tasks are recurrent in remote-sensing and medical imaging. A common approach is to solve the misalignment problem independently of the segmentation problem, by first aligning the noisy segmentation and image [3,8,9,28]. Most approaches require at least a small amount of high-quality annotations [7,28,32]. Closest to our work is [3], which introduces a consistency loss that applies pairs of random transformations to the misaligned labels and penalizes the network predictions when they are inconsistent with respect to the

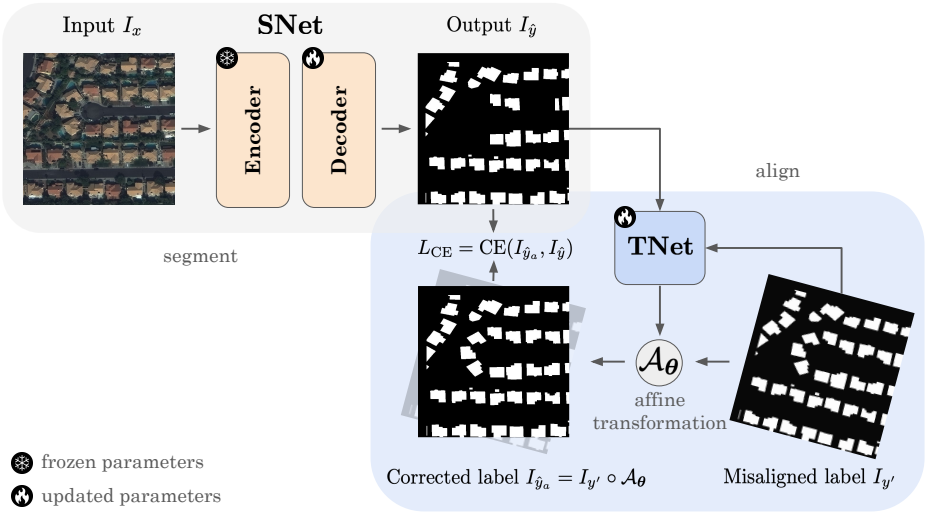

Fig. 2: Illustration of our Align and Segment (AnS) methodology. To directly learn semantic segmentation from misaligned labels, we introduce TNet, a lightweight module that learns to correct misaligned label maps while learning segmentation. This model-agnostic approach can be integrated in a plug-and-play manner with any segmentation network (SNet), as TNet directly operates on semantic segmentation maps.

applied transformations. However, without additional labels, the neural network can predict any biased set of predictions as long as they are relatively consistent with each other (e.g., all predictions may be erroneously shifted by the same constant offset). Our approach improves on this, as we predict transformations based on pairs of segmentation maps, which allows us to enforce absolute consistency with respect to a known random transformation. In [9], a multi-round procedure is proposed where, in each round, a model is trained to predict random transformations applied to the misaligned segmentation maps. This model is used to align the segmentation maps, which are then used to re-train the model in the next round. This method cannot work in the presence of biased transformations as the network is trained to predict an unbiased distribution of transformations. Although several studies also consider learning the segmentation map [7, 9, 14], none of them have attempted to jointly learn alignment and segmentation with the amount of shifts commonly found in remote sensing data. However, small translation and rotation errors were addressed by [7, 9, 32].

3 Methodology

We assume in the following that images and segmentation masks are defined in homogeneous coordinates and w.l.o.g. consider single channel images and binary segmentation problems, that is, we consider images as continuous functions of the form $I : \mathbb{P}^2 \to \mathbb{R}$.

3.1 Model architecture

In our framework, the inputs are given by an input image I_x and a corresponding misaligned segmentation mask $I_{y'}$. We assume that $I_{y'}$ is related to the aligned (unknown) segmentation mask I_y by an affine transformation $\mathcal{A}_{\theta^*}$ of the form

$$\mathcal{A}_\theta = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) & t_x \\ \sin(\alpha) & \cos(\alpha) & t_y \\ 0 & 0 & 1 \end{bmatrix} \quad , \tag{1}$$

where $\theta = (\alpha, t_x, t_y)$ are the three degrees of freedom of the transformation. With this, we have $I_{y'} = I_y \circ \mathcal{A}_{\theta^*}$. To predict this transformation, we employ a pair of networks, a TNet to predict the transformation between the image and the misaligned label mask, and an SNet to predict the semantic segmentation map from the image, see Fig. 2.

SNet. The segmentation network is a canonical deep neural network for semantic segmentation, mapping input images $I_x : \mathbb{P}^2 \to \mathbb{R}$ to probabilistic segmentation masks $I_{\hat{y}} : \mathbb{P}^2 \to [0, 1]$.

TNet. For the transformation network, our approach is inspired by the Spatial Transformer Network (STN, [12]) architecture introduced for classification[3], see the lower part of Fig. 2. In our AnS approach, the TNet takes as input the misaligned segmentation mask $I_{y'}$ and the segmentation mask $I_{\hat{y}}$, predicted by the SNet. Given $I_{y'}$ and $I_{\hat{y}}$, the TNet predicts the parameters of the affine transformation $\mathcal{A}_{\hat{\theta}}$. This predicted affine matrix is applied to $I_{y'}$ to yield the realigned segmentation mask $I_{\hat{y}_a} = I_{y'} \circ \mathcal{A}_{\hat{\theta}} = I_y \circ \mathcal{A}_{\theta^*} \circ \mathcal{A}_{\hat{\theta}}$.

To constrain the possible transformations, we apply an activation function $a(x) = c \tanh(x)$ to the network outputs. The scaling factor c limits the maximum alignment we can predict with the TNet, and can be based on the expected misalignment. The architecture can be trained via the cross entropy-loss

$$L_{\text{CE}} = \text{CE}(I_{y'} \circ \mathcal{A}_{\hat{\theta}}, I_{\hat{y}}) = \text{CE}(I_{\hat{y}_a}, I_{\hat{y}}) \ . \tag{2}$$

In practice, when dealing with rasterized rectangular images, the transformation $\mathcal{A}_{\hat{\theta}}$ leads to non-overlapping areas at the borders, which are excluded from the loss computation.

3.2 Handling transformation bias

The naive approach of minimizing L_{CE} fails, as the label map $I_{\hat{y}}$ predicted by the SNet can itself be translated with respect to I_x. Counterintuitively, this is the expected outcome. At the start of optimization, the TNet cannot predict the correct $\mathcal{A}_{\hat{\theta}} \approx A_{\theta^*}^{-1}$, and will likely predict a transformation close to the identity. Thus, at the start of optimization, we have $I_{\hat{y}} \approx I_{y'}$ and thus, the loss

[3] The name STN is *not* related to transformer networks based on attention [25].

of the SNet is given by comparing its prediction to the unaligned annotations. As a result, in the presence of bias in the label alignments, the SNet is driven towards predictions that are transformed to compensate for the misalignment. Since the cross-entropy can be potentially minimized to zero for the transformed maps, this is a stable solution, and thus the TNet will not learn the underlying translation.

We therefore developed two countermeasures: first, a consistency loss is added as an additional regularizer, which enables the TNet to move away from this unwanted local optimum. Second, data augmentations make it more difficult for the SNet to learn the biases in the dataset.

Consistency loss. To compute the consistency loss, we draw a random transformation $\mathcal{A}_{\theta_2}$ and apply it to the misaligned label map $I_{y'}$ to obtain $I_{y'} \circ \mathcal{A}_{\theta_2}$ similar to AutoCorrect [3]. Then we apply the TNet again to the input pair $(I_{y'} \circ \mathcal{A}_{\theta_2}, I_{y'})$ and predict $\mathcal{A}_{\hat{\theta}_2}$, which should ideally fulfill $\mathcal{A}_{\hat{\theta}_2} \approx \mathcal{A}_{\theta_2}^{-1}$. Since in this case, we know the optimal output of the TNet, we can measure both the cross-entropy of the re-aligned label-maps as well as the distance between $\mathcal{A}_{\hat{\theta}_2}$ and $\mathcal{A}_{\theta_2}^{-1}$. Most importantly, this prediction is independent of the segmentation map predicted by the SNet, allowing the TNet to move away from the unwanted local optimum. Moreover, it provides a learning signal early during training when the SNet has not developed a proper segmentation map yet.

The deviation of the predicted $\mathcal{A}_{\hat{\theta}_2}$ and the optimal $\mathcal{A}_{\theta_2}^{-1}$ is measured by the mean squared error (MSE) between the transformation matrix elements and the intersection over union (IoU) between the rasterized $I_{y_2'} \circ \mathcal{A}_{\hat{\theta}_2}$ and $I_{y'}$:

$$L_{\text{Con}} = \lambda \text{MSE}(\mathcal{A}_{\theta_2}^{-1}, \mathcal{A}_{\hat{\theta}_2}) + \text{IoU}(I_{y'}, I_{y'} \circ (\mathcal{A}_{\theta_2} \circ \mathcal{A}_{\hat{\theta}_2})) \tag{3}$$

The IoU loss emphasizes the overlapping regions between the two building masks, providing strong gradients when there is some overlap, while the MSE loss complements it by ensuring gradient flow even when there is little or no overlap between the two masks given to the TNet. As for the cross-entropy loss, non-overlapping regions are not considered in the IoU computation.

Augmentation. Even with the above loss functions, there is a risk that the segmentation model remains misaligned, as the required changes in the model parameters can be too large and the optimisation can get trapped in a local optimum. To overcome this issue, we applied strong augmentations both to the input image and misaligned mask to counteract the bias in the dataset. We do this by applying simple geometric augmentations such as vertical and horizontal flips, and rotations (90°, 180° and 270°) to the image pairs. For example, a horizontal flip of an image pair with an existing misalignment of 50 pixel translation on the x-axis changes to -50 pixel translation on the x-axis on the flipped image. In expectation, this removes the translation bias in the dataset.

Evaluation. We evaluate the performance using two variants of Intersection over Union (IoU) using 'golden' semantic reference maps I_y that are aligned with the input image I_x. Firstly, IoU_{seg} measures the performance of the segmentation network by comparing $I_{\hat{y}}$ and I_y:

$$\text{IoU}_{\text{seg}} = \text{IoU}(I_{\hat{y}}, I_y) \tag{4}$$

Secondly, $\text{IoU}_{\text{align}}$ measures the performance of the alignment network. It transforms the reference mask with a known transformation $\mathcal{A}_\theta^{-1}$ and measures how well the TNet can learn the inverse transformation:

$$\text{IoU}_{\text{align}} = \text{IoU}(I_y, I_y \circ \mathcal{A}_\theta^{-1} \circ \mathcal{A}_{\hat{\theta}}) \tag{5}$$

Errors in $\text{IoU}_{\text{align}}$ indicate erroneous estimates of the transformation parameters given a perfect segmentation.

Lastly, since there is a risk of the segmentation model learning misaligned labels directly, we are interested in evaluating if the modules learn the expected behavior, i.e. separating the alignment and segmentation. Therefore, we also report $\text{IoU}_{\text{learn}}$ computed by comparing $I_{\hat{y}_a}$ and $I_{\hat{y}}$:

$$\text{IoU}_{\text{learn}} = \text{IoU}(I_{y'} \circ \mathcal{A}_{\hat{\theta}}, I_{\hat{y}}) = \text{IoU}(I_{\hat{y}_a}, I_{\hat{y}}) \tag{6}$$

A high $\text{IoU}_{\text{learn}}$ but low IoU_{seg} and/or low $\text{IoU}_{\text{align}}$ indicates that the segmentation model learned to predict the misaligned labels and that the disentanglement of alignment and segmentation failed.

4 Dataset preparation

We studied building segmentation in remote sensing imagery from various cities across the world (see Tab. 1). To quantitatively study both random and systematic misalignment, we considered labeled data from Las Vegas, Paris, and Khartoum. We created two datasets per city from SpaceNet 2 [6], which contains WorldView-3 satellite imagery with 30 cm ground sampling distance and respective building footprints. The first is the biased dataset, $\mathcal{D}_{\text{bias}}$, where all masks are systematically translated by 50 pixels along the x-direction. As demonstrated in Fig. 1, label misalignment in geospatial applications is dominated by systematic noise. Learning from systematically misaligned data is particularly difficult, as a segmentation model can simply learn this misalignment, and, as a result, the input and the output map remain (geospatially) misaligned. The second dataset, $\mathcal{D}_{\text{uni}}$, was generated using randomly sampled affine transformations. If not further specified, we sampled transformations within the range $[-50, 50]$ pixels for t_x and t_y and $[-4.5, 4.5]$ degrees of α. This dataset spans a wide range of misalignment levels, from minimal to large geometric deviations. Correcting misalignment from $\mathcal{D}_{\text{uni}}$ is also challenging when the model is not sensitive to small misalignments. Fig. 3 shows example images from both datasets.

The synthetic data was generated for the first three cities in the table. The original image size is 650×650 pixels, which we split into patches of 320×320

Table 1: Dataset overview. Our synthetic datasets were constructed from the labels from the SpaceNet 2 dataset [6]. The first real-world dataset uses actual misaligned data from OpenStreetMap and imagery from the SpaceNet 5 dataset [6]. The second one, ReBO data [14] supplies both real misaligned and golden labels. Distributions of building sizes are given in the supplementary materials.

Cities	Source	Buildings	Mean (m^2)	Max (m^2)	Patches	Golden	Real
Las Vegas	SpaceNet 2	108,943	206	28,702	13,280	✓	✗
Paris	SpaceNet 2	16,478	121	14,775	2,111	✓	✗
Khartoum	SpaceNet 2	25,113	240	17,841	3,077	✓	✗
San Juan	SpaceNet 5 & OSM	284,042	218	113,909	6,726	✗	✓
41 cities	ReBO	179,265	354	54,808	5473	✓	✓

pixels (more details about the data generation are provided in the supplementary material). After applying the systematic bias and uniformly sampled transformations to all image pairs, we split all pairs from each city, including real-world data, into training, validation, and test sets randomly, with 80%, 10% and 10%, respectively (see Tab. 1).

To demonstrate that our method works on OpenStreetMap [21] building data, we collected WorldView-3 imagery from a fourth city, San Juan, through the SpaceNet 5 [6] dataset. These data were released without building footprints, and we downloaded corresponding footprints from OpenStreetMap and prepared them for alignment and segmentation with our methodology.

In addition, we evaluated our method on ReBO data [14], $\mathcal{D}_{\text{ReBO}}$, a real-world dataset that includes real misaligned and golden labels. The dataset consists of imagery and polygon labels for each building, including footprints, roof, and OSM polygons. The OSM polygons are real-world labels, and the remaining two are manually corrected labels delineating building footprints and roofs, respectively. The imagery consists of both off- and near-nadir images. We use only roof labels as golden labels in our experiments because footprint labels fuse with facade pixels in off-nadir images. While the other datasets were generated for individual cities, $\mathcal{D}_{\text{ReBO}}$ comprises patches from 41 cities worldwide. The dataset was released with a training and test split. We used the training set for both training and validation and reported final results on the test set. The spatial resolution is 50 cm, and each patch in the data has a size of 512×512 pixels. See Tab. 1 for more details about this dataset.

5 Experiments

In our experiments, the TNet used the ViT-Small [5] architecture and the SNet was based on a U-Net [22], where the encoder was replaced by ConvNeXt-Tiny [17]. In all our experiments, we used ConvNeXt-Tiny with DINOv3 [23] pretrained weights and froze the encoder. We also experimented with alternative encoders in SNet (see supplementary materials).

We conducted five sets of experiments:

1. **Regularization, augmentation, and end-to-end training.** We evaluated the effects of the transformation bias mitigation strategies. We either trained with L_{CE} or $L_{\text{CE}} + L_{\text{Con}}$, and either enabled or disabled the large-scale image augmentations, resulting in four different settings. The consistency loss (3) does depend on the SNet output. This suggests to pretrain the TNet before training the SNet training, and we evaluated such a two-stage process. All of these experiments were conducted on the Las-Vegas dataset, the two-stage learning was additionally evaluated on $\mathcal{D}_{\text{ReBO}}$.
2. **Robustness.** To test the robustness of our method against the magnitude of transformations, we varied the magnitude of the applied transformations in $\mathcal{D}_{\text{uni}}$ and $\mathcal{D}_{\text{bias}}$ between 10–100 pixels on the Las-Vegas dataset. For each setting, we re-trained the model and evaluated IoU (see Sec. 6).
3. **Baseline comparison.** We compared our method with five baseline models. These include, two supervised methods, such as MapRepair (MR) [32], and Alignment Correction Network (ACN) [7], and three unsupervised methods, such as Map Alignment (MA) [9], AutoCorrect (AC) [3], and Spatial Correction (SC) [27]. We tried our best to implement these methods for a fair comparison; e.g., we used our TNet architecture in both MR and AC. While some of these baselines were trained to fit only the training set, we used standard data splits such as train, validation, and test sets for all our experiments. SC and MA correct labels by training a model in multiple rounds. SC uses golden labels from the validation dataset and is consequently not a fully unsupervised method. Both MA and our method do not need any golden labels. A crucial hyperparameter of prior work is the number of *rounds* after which the model corrects the labels.
4. **Real-world dataset.** We performed an experiment on the $\mathcal{D}_{\text{ReBO}}$ dataset [14] that provides real misalignments and golden labels for 41 cities. While they released the patches with the size of 512×512 pixels, we predicted a single affine transformation for each patch. We used this dataset to train and compare our method with all the baselines.
5. **Qualitative analysis.** Our method was applied to a real-world dataset with building footprints from OpenStreetMap paired with WorldView-3 imagery.

Details of the training process. In all experiments and for all models, we used the AdamW [18] optimizer with a learning rate of 0.00001 for training. The model was trained for 300 epochs with a batch size of 48. After training, we selected the best model with the weights of the epoch with highest $\text{IoU}_{\text{learn}}$ score on the validation set. Since golden labels I_y are unavailable in real-world scenarios, $\text{IoU}_{\text{align}}$ and IoU_{seg} were used only for monitoring and reporting purposes and not for model selection.

Additional hyperparameters. We set $c = 0.35$ for scaling the tanh activation, which constrains the translations between $[-112, 112]$ pixels along both axes and the rotation range to be within $[-20.06, 20.06]$ degrees around the image

Table 2: Regularization and augmentations. We study the effect of the regularization loss L_{Con} and augmentations. The results show the performance on Las Vegas with the frozen ConvNeXt-Tiny encoder with DINOv3 weights. Both components lead to complementary improvements, with largest improvements on $\mathcal{D}_{\text{bias}}$.

			$\mathcal{D}_{\text{uni}}$			$\mathcal{D}_{\text{bias}}$		
L_{CE}	L_{Con}	Aug.	IoU_{seg}	$\text{IoU}_{\text{align}}$	$\text{IoU}_{\text{learn}}$	IoU_{seg}	$\text{IoU}_{\text{align}}$	$\text{IoU}_{\text{learn}}$
✓	-	-	0.71	0.77	0.72	0.39	0.41	0.79
✓	✓	-	0.74	0.79	0.72	0.41	0.43	0.78
✓	-	✓	0.66	0.69	0.79	0.72	0.76	0.82
✓	✓	✓	0.79	0.84	0.79	0.78	0.88	0.82

center. We further set $\lambda = 100$ to scale the two loss terms in Eq. (3) to a similar order of magnitude. In the experiments using augmentation, all five geometric augmentations were applied with an equal probability of 50% to each batch. All experiments were carried out on AMD MI250X (64GB) GPUs.

6 Results

Regularization, augmentation, and end-to-end training. The results for this experiment are given in Tab. 2. The results show that both of our regularization techniques improved IoU_{seg} and $\text{IoU}_{\text{align}}$ for $\mathcal{D}_{\text{uni}}$ and $\mathcal{D}_{\text{bias}}$ compared to the baseline of only using L_{CE}. Again, the effect was more pronounced for $\mathcal{D}_{\text{bias}}$, where a sharp decay in both IoU_{seg} and $\text{IoU}_{\text{align}}$ could be observed in the baseline, while $\text{IoU}_{\text{learn}}$ did not vary at all, showing that the SNet was still able to predict the misaligned segmentation maps correctly. In our experiments, adding the augmentations when dealing with the biased $\mathcal{D}_{\text{bias}}$ gave the largest improvement, whereas adding only the augmentation to $\mathcal{D}_{\text{uni}}$ decreased the accuracy.

We compared our end-to-end learning approach with sequential training. We first trained the TNet alone with L_{Con}, then froze it and trained the whole architecture with L_{CE}. This strategy failed to produce the desired results. On $\mathcal{D}_{\text{uni}}$, $\mathcal{D}_{\text{bias}}$, and $\mathcal{D}_{\text{ReBO}}$, the two-stage training only achieved an IoU_{seg} of 0.48, 0.44, and 0.38, respectively, and an $\text{IoU}_{\text{align}}$ of 0.31, 0.42, and 0.27, respectively (cf. Tab. 3). Learning the whole system with pretrained and frozen TNet seems to make it difficult to provide guiding transformations in early stages of SNet learning. The reason is the TNet pretraining was based on misaligned but otherwise (almost) perfect masks (we added pixel noise) and the SNet produces very noisy masks in the beginning. There might be even better ways to combine TNet and SNet optimization than our simultaneous training, but we leave this question to future work.

Baseline comparison on synthetic and real datasets. The results of our performance comparison are summarized in Tab. 3. Our method outperformed all

Table 3: Comparison to baseline methods. For each city, the baseline methods MapRepair (MR) [32], Alignment Correction Network (ACN) [7], Spatial Correction (SC) [27], Map Alignment (MA) [9], and AutoCorrect (AC) [3] are compared to our *AnS* method trained on both $\mathcal{D}_{\text{uni}}$ and $\mathcal{D}_{\text{bias}}$ and evaluated on the respective test set. For every city, the first row indicates the IoU metrics of the *initial overlap* of the misaligned labels without any correction. The results on the test set of $\mathcal{D}_{\text{ReBO}}$ are added next to Las Vegas city as additional two columns. (*Denotes supervised baselines.)

		$\mathcal{D}_{\text{uni}}$		$\mathcal{D}_{\text{bias}}$		$\mathcal{D}_{\text{ReBO}}$	
City	Method	IoU_{seg}	$\text{IoU}_{\text{align}}$	IoU_{seg}	$\text{IoU}_{\text{align}}$	IoU_{seg}	$\text{IoU}_{\text{align}}$
Las Vegas	Initial overlap	0.53	0.53	0.46	0.46	0.54	0.54
	MR [32]*	–	0.77	–	0.77	–	0.76
	ACN [7]*	0.94	–	0.93		0.87	–
	SC [27]	0.66	–	0.47	–	0.49	–
	MA [9]	0.55	0.70	0.49	0.47	0.55	0.57
	AC [3]	–	0.65	–	0.46	–	0.60
	AnS (ours)	0.79	0.84	0.78	0.88	0.62	0.74
Paris	Initial overlap	0.47	0.47	0.40	0.40		
	MR [32]*	–	0.57	–	0.68		
	ACN [7]*	0.88	–	0.87	–		
	SC [27]	0.48	–	0.40	–		
	MA [9]	0.49	0.51	0.42	0.37		
	AC [3]	–	0.54	–	0.36		
	AnS (ours)	0.56	0.67	0.52	0.65		
Khartoum	Initial overlap	0.58	0.58	0.50	0.50		
	MR [32]*	–	0.71	–	0.75		
	ACN [7]*	0.91	–	0.90	–		
	SC [27]	0.60	–	0.47	–		
	MA [9]	0.60	0.61	0.53	0.46		
	AC [3]	–	0.54	–	0.43		
	AnS (ours)	0.66	0.77	0.63	0.71		

unsupervised baselines (SC, MA, AC) in both IoU_{seg} and $\text{IoU}_{\text{align}}$, on both synthetic and real datasets. As expected, both supervised methods (MR, ACN) performed well compared to the unsupervised baselines and our method. Still, our method is competitive with MR, outperforming on $\mathcal{D}_{\text{uni}}$ and $\mathcal{D}_{\text{bias}}$ in Las Vegas, and on $\mathcal{D}_{\text{uni}}$ in Paris and Khartoum. Note that ACN uses misaligned labels as an input, which means that it required misaligned labels during inference, while no other models require misaligned labels for segmentation. The model performed best on Las-Vegas (which could be an artifact of selecting our architecture on this model), however, it also performed better for Khartoum and Paris, which showed a much smaller initial overlap under our transformations

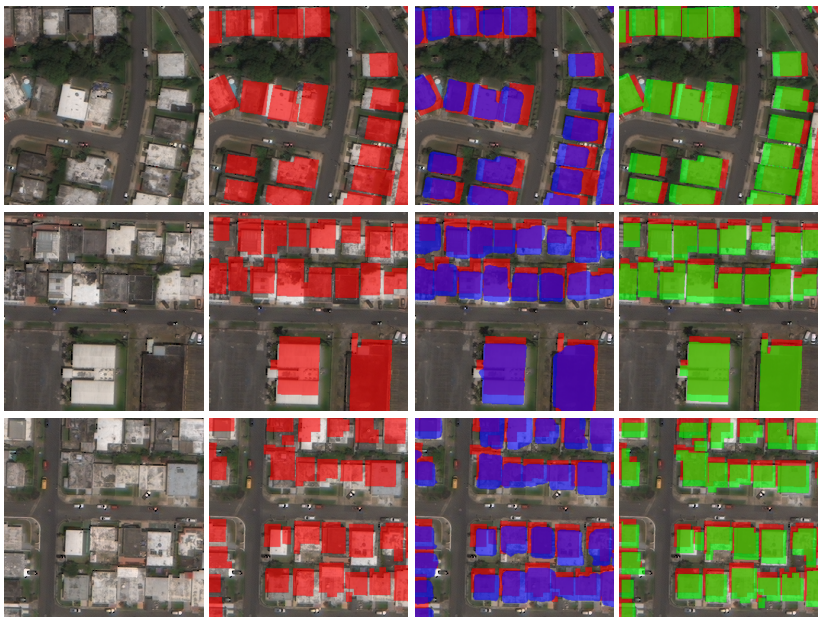

Fig. 3: Qualitative results learned on OpenStreetMap data. We show qualitative examples correcting real-world misaligned labels from OpenStreetMap in San Juan using our method. The first column shows the RGB image (I_x). All other columns display the misaligned building mask ($I_{y'}$) on top of I_x for the reference. The predicted segmentation mask ($I_{\hat{y}}$) and aligned mask ($I_{\hat{y}_a}$) are shown in the third and fourth column, respectively. While the output of the semantic segmentation network $I_{\hat{y}}$ has blurry edges and corners (3rd column), the corrected segmentation map $I_{\hat{y}_a}$ using the estimated affine transformation preserves topology and shapes (last column).

than the other two cities. Note that for Paris, $\text{IoU}_{\text{align}}$ was consistently large for our method, showing that the problem is likely a result of difficulties in predicting the segmentation and a small training dataset, while the transformations were still predicted correctly.

Qualitative analysis on OpenStreetMap data. The results of the qualitative analysis on the real application of our method are given in Fig. 3. Since we do not have golden labels here, we can only estimate the overlap between predicted segmentation map and aligned segmentation map, $\text{IoU}_{\text{learn}}$, and we obtained $\text{IoU}_{\text{learn}}$=0.70. Three randomly selected examples of the model predictions are given in Fig. 3, and more examples are given in the supplementary material. As can be seen, the predicted transformations correct the misalignment of the label map well. However, the produced label maps are blurry and often do not represent the actual shapes of the houses. This is likely due to the presence of label-errors as we have not included any robust loss components for missing house annotations.

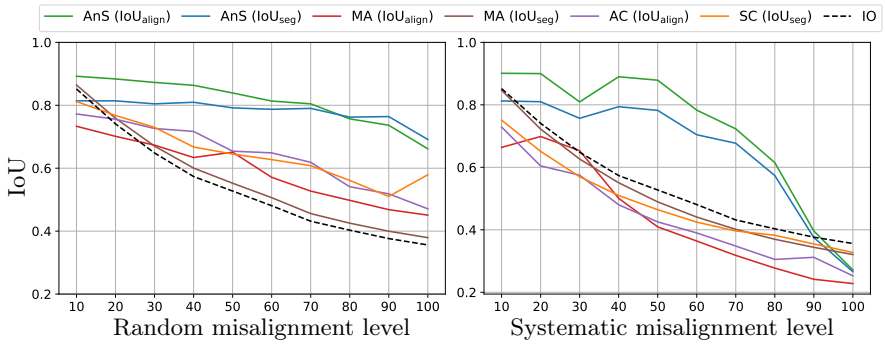

Fig. 4: Robustness to different noise levels. Sensitivity analysis of training the model with increasing noise-levels on the random misalignment dataset $\mathcal{D}_{\text{uni}}$ and systematic misalignment dataset $\mathcal{D}_{\text{bias}}$.

Robustness. Fig. 4 shows the results of our robustness experiment with different noise levels. We plot IoU_{seg} and $\text{IoU}_{\text{align}}$ of our method and all unsupervised baselines, and the IoU of the misaligned transformations with the golden labels (initial overlap, IO) for comparison. Compared to the initial overlap, all baselines performed slightly better on $\mathcal{D}_{\text{uni}}$ but did not improve over the initial overlap on $\mathcal{D}_{\text{bias}}$. The high IoU_{seg} from our method shows that our model is robust against the magnitude of the applied transformation until 100 pixels in $\mathcal{D}_{\text{uni}}$ and about 70 pixels in $\mathcal{D}_{\text{bias}}$. Then the IoU_{seg} starts to drop, which is reflected by a simultaneous drop in $\text{IoU}_{\text{align}}$.

7 Discussion

Our Align-and-Segment (AnS) framework has been designed to learn semantic segmentation models from systematically misaligned labels without requiring any access to golden labels. On all three cities in the artificial datasets as well as on the real-world data with and without golden labels, our approach demonstrated that jointly estimating label alignment and segmentation is not only feasible but can substantially improve segmentation performance.

Standard segmentation networks readily overfit to misaligned labels. As our experiments show, simply adding a spatial transformer module, referred to as TNet, does not solve the problem, as it is easily bypassed by the segmentation network. However, our proposed loss term forces the TNet to learn and perform consistent transformations, and the TNet indeed learns to accurately estimate the underlying alignment transformations. As expected, random misalignments were easier to address than systematic misalignments. However, we showed that data augmentation in the form of rotating and flipping training images together with their misaligned segmentation masks significantly improves the performance in this setting.

We identified several issues with the baseline methods considered in our experiments. For example, SC and MA required significant training time due to

iterative label correction, incurring computational overhead. In our experiments, we trained these models for at least three rounds as recommended [9, 27], which means the computational cost is three times higher than single-round approaches like AC and ours. MA assigns a single random transformation to the misaligned mask in each round, i.e., the same transformations are used across multiple epochs, whereas our approach samples a new transformation in each epoch. Resampling transformations every epoch reduces the risk of learning the misalignment. While AC also uses a similar sampling approach with a consistency loss, there is an important difference: In AC, the transformation network operates on two different modalities, RGB images and segmentation masks, whereas our TNet uses two segmentation masks as input.

Limitations. A limitation of our work is the use of a generic decoder for the building footprint segmentation. We neither optimized the decoder for the task at hand nor used any post-processing. Specialized methods for building segmentation would surely improve the delineation of the building boundaries (e.g., [10, 11]).

Furthermore, the segmentation is hampered by label noise beyond misalignments, and adding, for example, mechanisms for handling missing labels would improve the training. The current model is restricted to learning affine transformations at the scale of the input patch, but the approach is general enough to allow for more complex transformations.

8 Conclusion

Global building footprint data are key for monitoring population densities and managing cities on a large scale. Consequently, several organizations now provide large-scale building-footprint datasets, including Google's Open Buildings [24] and Microsoft building footprints [4], supplementing globally curated contributions available through OpenStreetMap [21]. Such existing building data could be automatically combined with satellite imagery to construct training datasets to learn semantic segmentation models, which would allow to generate inexpensive high-quality building-footprint data with extensive coverage and high temporal resolution. However, progress is hampered by the misalignment between independent imagery and existing footprint data.

In this work, we propose *AnS*, a model-agnostic methodology that can be combined with any semantic segmentation model to directly learn the alignment and the segmentation from misaligned labels. We empirically evaluated our end-to-end approach and demonstrated that regularization and data augmentation are key to disentangle segmentation and alignment and to avoid undesired shortcut learning of misaligned labels. On synthetic data, we demonstrate that AnS is robust to high noise levels. Furthermore, we show that our approach can directly learn from real-world misaligned data (e.g., labels from OpenStreetMap) without using any golden labels.

Acknowledgements

We acknowledge support through the project *Risk-assessment of Vector-borne Diseases Based on Deep Learning and Remote Sensing* funded by the Novo Nordisk Foundation (grant number NNF21OC0069116). AK and CI acknowledge additional support by the *Center for Remote Sensing and Deep Learning of Global Tree Resources (TreeSense)* funded by the Danish National Research Foundation (grant number DNRF192). NL and CI acknowledge support by the *Global Wetland Center* (GWC) funded by the Novo Nordisk Foundation (grant number NNF23OC0081089). This work was supported in part by the Pioneer Centre for AI, DNRF grant number P1. We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC (Finland) and the LUMI consortium through a EuroHPC Regular Access call.

References

1. Bettencourt, L.M.A., Marchio, N.: Infrastructure deficits and informal settlements in sub-Saharan Africa. Nature **645**(8080), 399–406 (2025)
2. Bradbury, K., Brigman, B., Collins, L., Johnson, T., Lin, S., Newell, R., Park, S., Suresh, S., Wiesner, H., Xi, Y.: Aerial imagery object identification dataset for building and road detection, and building height estimation. figshare (2016)
3. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: AutoCorrect: Deep inductive alignment of noisy geometric annotations. In: British Machine Vision Conference (2019)
4. Corporation, M.: Microsoft global ML building footprints. `https://planetarycomputer.microsoft.com/dataset/ms-buildings` (2023), accessed: 2025-11-15
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
6. Etten, A.V., Lindenbaum, D., Bacastow, T.M.: SpaceNet: A remote sensing dataset and challenge series (2019)
7. Fobi, S., Conlon, T., Taneja, J., Modi, V.: Learning to segment from misaligned and partial labels. In: ACM Conference on Computing and Sustainable Societies (SIGCAS). p. 286–290 (2020)
8. Girard, N., Charpiat, G., Tarabalka, Y.: Aligning and updating cadaster maps with aerial images by multi-task, multi-resolution deep learning. In: Asian Conference on Computer Vision (ACCV). pp. 675–690 (2019)
9. Girard, N., Charpiat, G., Tarabalka, Y.: Noisy supervision for correcting misaligned cadaster maps without perfect ground truth data. In: International Geoscience and Remote Sensing Symposium (IGARSS). pp. 10103–10106 (2019)
10. Girard, N., Smirnov, D., Solomon, J., Tarabalka, Y.: Polygonal building extraction by frame field learning. In: Computer Vision and Pattern Recognition (CVPR). pp. 5891–5900 (2021)
11. Guthula, V.B., Oehmcke, S., Chilaule, R., Zhang, H., Lang, N., Kariryaa, A., Mottelson, J., Igel, C.: Drone imagery for roof detection, classification, and segmentation to support mosquito-borne disease risk assessment: The Nacala-roof-material dataset. Science of Remote Sensing p. 100306 (2025)

12. Jaderberg, M., Simonyan, K., Zisserman, A., kavukcuoglu, k.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 28 (2015)
13. Kun, Y., Jianxin, W.: Probabilistic end-to-end noise correction for learning with noisy labels. In: Computer Vision and Pattern Recognition (CVPR) (2019)
14. Li, K., Weng, X., Deng, Y., Meng, Y., Pang, C., Xia, G.S., Zhao, X.: DragOSM: Extract building roofs and footprints from aerial images by aligning historical labels (2025)
15. Liu, S., Liu, K., Zhu, W., Shen, Y., Fernandez-Granda, C.: Adaptive early-learning correction for segmentation from noisy annotations. In: Computer Vision and Pattern Recognition (CVPR). pp. 2606–2616 (2022)
16. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33 (2020)
17. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Computer Vision and Pattern Recognition (CVPR). pp. 11976–11986 (2022)
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
19. Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P.: Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In: International Geoscience and Remote Sensing Symposium (IGARSS). pp. 3226–3229 (2017)
20. Maiti, A., Oude Elberink, S., Vosselman, G.: Effect of label noise in semantic segmentation of high resolution aerial images and height data. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **2**, 275–282 (2022)
21. OpenStreetMap contributors: Planet dump retrieved from `https://planet.osm.org`. `https://www.openstreetmap.org` (2017)
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 234–241. Springer (2015)
23. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025)
24. Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y.S.E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M., Quinn, J.: Continental-scale building detection from high resolution satellite imagery (2021)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Processing Systems (NeurIPS). vol. 30 (2017)
26. Yao, J., Zhang, Y., Zheng, S., Goswami, M., Prasanna, P., Chen, C.: Learning to segment from noisy annotations: A spatial correction approach. In: International Conference on Learning Representations (ICLR) (2023)
27. Yao, J., Zhang, Y., Zheng, S., Goswami, M., Prasanna, P., Chen, C.: Learning to segment from noisy annotations: A spatial correction approach. In: International Conference on Learning Representations (ICLR) (2023)

28. Zampieri, A., Charpiat, G., Girard, N., Tarabalka, Y.: Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing. In: European Conference on Computer Vision (ECCV) (2018)
29. Zhang, M., Gao, J., Lyu, Z., Zhao, W., Wang, Q., Ding, W., Wang, S., Li, Z., Cui, S.: Characterizing label errors: Confident learning for noisy-labeled image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 721–730 (2020)
30. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 8792–8802 (2018)
31. Zhu, H., Shi, J., Wu, J.: Pick-and-Learn: Automatic quality evaluation for noisy-labeled image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 576–584 (2019)
32. Zorzi, S., Bittner, K., Fraundorfer, F.: MapRepair: Deep cadastre maps alignment and temporal inconsistencies fix in satellite images. In: International Geoscience and Remote Sensing Symposium (IGARSS). pp. 1829–1832 (2020)