Self-Supervised Feature Learning for Semantic Segmentation of Overhead Imagery

Suriya Singh*[1]
Anil Batra*[1]
Guan Pang[2]
Lorenzo Torresani[3]
Saikat Basu[2]
Manohar Paluri[2]
C. V. Jawahar[1]

[1] IIIT Hyderabad
[2] Facebook
[3] Dartmouth College

Abstract

Overhead imageries play a crucial role in many applications such as urban planning, crop yield forecasting, mapping, and policy making. Semantic segmentation could enable automatic, efficient, and large-scale understanding of overhead imageries for these applications. However, semantic segmentation of overhead imageries is a challenging task, primarily due to the large domain gap from existing research in ground imageries, unavailability of large-scale dataset with pixel-level annotations, and inherent complexity in the task. Readily available vast amount of unlabeled overhead imageries share more common structures and patterns compared to the ground imageries, therefore, its large-scale analysis could benefit from unsupervised feature learning techniques.

In this work, we study various self-supervised feature learning techniques for semantic segmentation of overhead imageries. We choose image semantic inpainting as a self-supervised task [36] for our experiments due to its proximity to the semantic segmentation task. We (i) show that existing approaches are inefficient for semantic segmentation, (ii) propose architectural changes towards self-supervised learning for semantic segmentation, (iii) propose an adversarial training scheme for self-supervised learning by increasing the pretext task's difficulty gradually and show that it leads to learning better features, and (iv) propose a unified approach for overhead scene parsing, road network extraction, and land cover estimation. Our approach improves over training from scratch by more than 10% and ImageNet pre-trained network by more than 5% mIOU.

1 Introduction

Overhead imageries are images captured by imaging satellites, aeroplanes, drones, etc. They can be updated easily as well as frequently [29]. In contrast to ground imageries which are often captured with digital, portable, or surveillance cameras, overhead imageries present a unique and occlusion-free view of a large geographical area (see Figure 1 (a)). Due to this, they are extensively used for land cover classification [15, 24], scene parsing [1, 20, 35], road network extraction [3, 4, 28, 29, 30, 31, 32, 43], etc. However, the focus has been towards

* Equal Contributions.

Code available at: https://github.com/suriyasingh/Self-supervision-for-segmenting-overhead-imagery

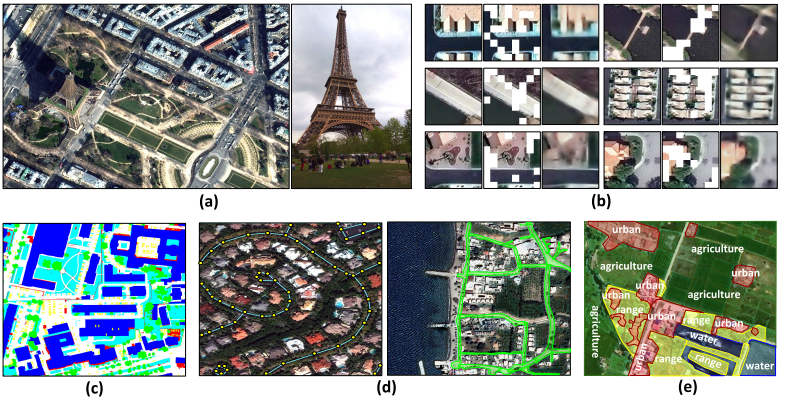


Figure 1: (a) There exist a large domain gap between ground and overhead imageries due to the perspective* . (b) Overhead imageries exhibit rich context information making it suitable for unsupervised feature learning with inpainting. Semantic segmentation of overhead imagery enables a variety of tasks: (c) scene parsing of a city, (d) road network extraction, and (e) land cover estimation.

a specific narrow area of application. In this work, we present a unified approach, based on semantic segmentation, towards a variety of overhead imagery tasks — (i) scene parsing of a city, (ii) road network extraction in urban and remote areas, and (iii) land cover estimation in diverse geographical terrains.

Overhead imagery captures a vast geographical area with diverse landscapes, objects and extreme variations in their count, size, and aspect-ratio. Moreover, significant diversity arises due to illumination, region's geography, weather conditions, etc. Undoubtedly, there is a large domain gap from existing research in ground imageries primarily due to the overhead perspective. Transfer learning between these domains is, therefore, unsuitable. Overhead view-point of objects and scenes are highly ambiguous and their annotations require domain expertise due to the uncommon appearance. Unavailability of large-scale dataset with pixel-level annotations for different overhead imagery tasks further limits the utility of current semantic segmentation techniques [2, 25, 26, 33].

In this work, we exploit the strong context information and spatial relationship present in overhead imageries to learn useful features at the pre-training stage with self-supervised technique. We employ semantic inpainting as the self-supervised task [36] (Figure 1 (b)) due to its proximity to the semantic segmentation task. We propose architectural changes (3.1) enabling the pre-training of encoder network which preserves the spatial context of features as well as the decoder network which learns to upsample the features with respect to the semantic boundary of entities, an essential ingredient for semantic segmentation.

Semantic inpainting as self-supervised task leads to learning useful features only when the region filling task is non-trivial. Curated object-centric datasets (ImageNet [7], Pascal VOC [9], etc.) are inherently diverse and objects occupy a significant portion of the image. Erasing fixed or random regions from object centric images, therefore, is adequately difficult. On the contrary, overhead imageries with much wider world-view lacks specific subject in the images. To ensure the pretext task's difficulty, instead of inpainting random regions [36] (Figure 1 (b) left), we propose to inpaint difficult and semantically meaningful regions (Figure 1 (b) right) with an adversarial training scheme consisting of coach and inpainting networks (3.2). The coach network see an entire image and predicts an increasingly difficult

* Image source: http://visions-of-earth.com/satellite-image-of-paris-france-the-city-of-love-and-romance/

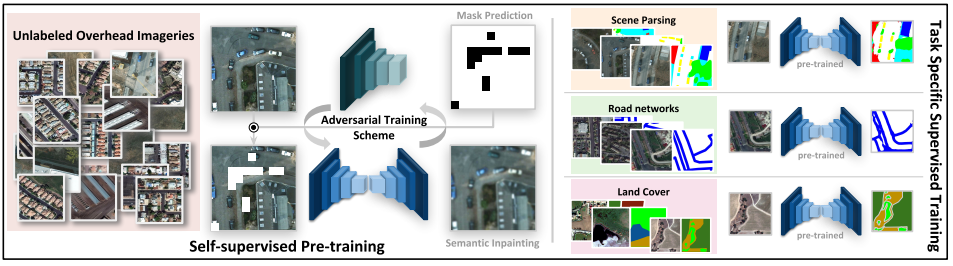


Figure 2: An overview of our approach: self-supervised pre-training (left) and task specific supervised training (right). We use semantic inpainting as self-supervised pre-training [36] to exploit the freely available large amount of unlabeled overhead imageries. To ensure the pretext task's difficulty we train the inpainting network with an adversarial mask prediction scheme. The pre-trained encoder-decoder inpainting network is then fine-tuned for a variety of overhead imagery tasks: scene parsing, road network extraction, and land cover estimation.

mask which is used to erase the corresponding regions of the image. The inpainting network then tries to fill-in the erased regions with the help of available contexts. At the end of the pre-training stage, the inpainting network learn to efficiently encode the available contexts and upsample the activation maps for overhead imageries. The pre-trained model is further used as initialization for different overhead imagery tasks. Figure 2 shows the overview of our approach.

Contributions

1. We show that existing self-supervised techniques focusing on the encoder network alone are inefficient for semantic segmentation. We propose architectural changes towards self-supervised pre-training of encoder as well as decoder networks.
2. We propose an adversarial training scheme for self-supervised learning by increasing the pretext task difficulty gradually and show that it leads to superior performance.
3. We also propose a unified segmentation based approach for scene parsing, road network extraction, and land cover estimation in overhead imageries. Our technique improves over training from scratch by more than 10% and ImageNet pre-trained network by more than 5% mIOU.

2 Related Works

Overhead Imagery Understanding The overhead imagery community, in the past, has mostly focused on specific task and application individually. The prominent tasks in this domain are land cover classification [15, 24], scene parsing [1, 20, 35], and road network extraction [3, 4, 28, 29, 30, 31, 32, 43]. Readers are suggested to see [46] for a comprehensive survey on recent developments in overhead imagery analysis. Unsupervised input reconstruction [30, 31], supervised pre-training on natural images [1], and data augmentations with balancing class population [20] have been explored to overcome the data scarcity. In contrast to these works, we perform pre-training with self-supervision from the same domain and show its efficacy in a unified semantic segmentation approach for scene parsing, road network estimation, and land cover estimation.

Unsupervised and Self-Supervised Feature Learning Deep learning models require a large amount of annotated data to train from scratch. RBMs [16], Autoencoders [17], and its

variants [37, 40, 41] have been popular choice for unsupervised pre-training where labeled data is scarce [30, 31]. Recently, self-supervised learning techniques [8, 11, 34, 36, 42, 45] using freely available *pseudo* labels have emerged as a superior technique due to stronger self-supervision. Doersch *et al.* [8] proposed to learn representations by predicting relative position of two patches in an image. Noroozi *et al.* [34] extended this idea further to train the network for solving jigsaw puzzles. Zhang *et al.* [45] proposed Split-Brain Autoencoders, two disjoint sub-networks each trained to predict the missing image channel(s). Pathak *et al.* [36] proposed Context Encoders to predict the contents of missing regions in the image using the available contexts. Note that, [8, 11, 34, 36, 45] focus on pre-training the encoder networks alone, and therefore, are inefficient for semantic segmentation. Furthermore, difficulty level of the hand-crafted self-supervised tasks are fixed across examples depending on the nature of the task itself. In this work, we propose an adversarial training scheme capable of generating increasingly difficult examples for pre-training based on content of the image.

Semantic Segmentation Recent semantic segmentation [2, 25, 26, 33] techniques rely on backbone model pre-trained on related task such as supervised image classification. Self-supervised pre-training have also shown promising results on popular benchmarks [36, 45]. In both supervised [2, 25, 26, 33] and self-supervised pre-training [36, 45], the decoder network is trained from scratch for semantic segmentation. In contrast to this, we propose pre-training of the encoder as well as the decoder networks with semantic inpainting task.

3 Method

3.1 Semantic Inpainting as Self-supervision

Image semantic inpainting refers to predicting the actual image x from its corrupted version $\hat{x}$. The inpainting model learns from the available contexts in $\hat{x}$ to reconstruct x. Pathak *et al.* [36] proposed semantic inpainting as self-supervision to learn visual representation of the image. In [36], a random binary mask M is generated for each image such that the pixels with corresponding mask value 0 are erased from the image, and 1 are kept intact.

$$\hat{x} = M \odot x \tag{1}$$

where $\odot$ is the element-wise product operation. The inpainting model F learns to inpaint images by minimizing the masked L_2 distance as reconstruction loss, $\mathcal{L}_{rec}$. We add additional loss term for context regions, $\mathcal{L}_{con}$, to allow the network to reconstruct the entire image and learn to upsample activation maps effectively.

$$L_{rec}(\hat{x}) = \frac{1}{\sum(1-M)} ||(1-M) \odot (x - F(M \odot x))||_2^2 \tag{2}$$

$$L_{con}(\hat{x}) = \frac{1}{\sum M} \qquad ||M \odot (x - F((1-M) \odot x))||_2^2 \tag{3}$$

The final loss $\mathcal{L}_{inpainting}$ is the weighted sum of reconstruction and context losses.

$$L_{inpainting} = \mathbf{w}_{rec}\mathcal{L}_{rec} + \mathbf{w}_{con}\mathcal{L}_{con} \tag{4}$$

Architectural Improvements We propose architectural changes to the semantic inpainting encoder-decoder architecture used in [36]. We use **(a)** a more powerful ResNet-18 [13]

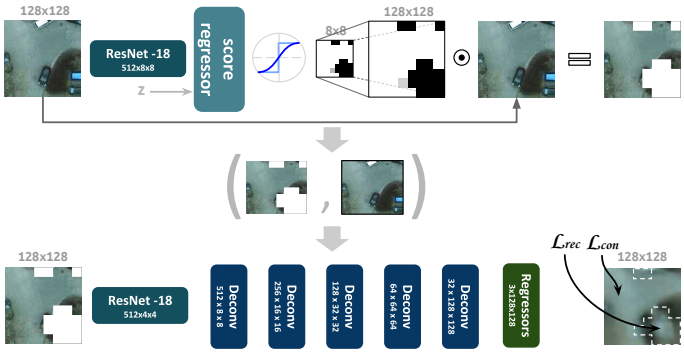


Figure 3: The coach network (top) take the image as input and outputs a semantically meaningful binary mask. The mask is used to erase parts of the image which then is used as input to the inpainting network (bottom). The inpainting network learns to encode visual representations as well as upsampling by trying to fill-in the image regions erased by the coach. The coach is trained with loss adversary to the inpainting network making it capable of generating increasingly difficult examples (see 3.2).

as the backbone encoder network, **(b)** do away with the channel-wise fully-connected bottleneck layer, and **(c)** exploit the pre-trained encoder as well as decoder for the segmentation task. ResNet [13], compared to AlexNet [23] equivalent used in [36], have the potential to learn better representations, is more efficient as well as easier and faster to train [14]. Furthermore, `BatchNorm` [18] helps in reducing the domain gap between semantic inpainting of corrupted images and semantic segmentation of natural images since the input to convolutional layers follow the same distribution during both stages.

Fully-connected bottleneck layer in an encoder-decoder network connects all spatial locations together, however, also results in losing the vital spatial context. Deep CNNs' (AlexNet [23], VGG [39], ResNet [13]) convolutional filters possess large enough field-of-view (FOV) to *see* the spatial extent of 195×195 pixels (or more) of input [27]. We use the input size 128×128 for inpainting which is well within reach of the encoder network's FOV. By **not** employing the fully-connected bottleneck layer in our architecture, the resulting network is fully convolutional, able to preserve the spatial context, and has fewer parameters.

Lastly, while learning to inpaint, the decoder network tries to push the low resolution feature up to the semantic boundary of the entities at input resolution. The decoder network learns non-linear weighted upsampling of the low resolution feature maps which we show is useful for the target segmentation task. To the best of our knowledge, ours is the first architecture that re-uses the encoder as well as the decoder network for the target task.

3.2 Coach Network

Pathak *et al.* [36] inpaints the image erased by a randomly generated binary mask. The mask dictates the regions used to learn context, the regions to inpaint, the difficulty of the task, and in turn the quality of the learned features. Overhead imageries with much wider world-view lacks specific subject in the images, therefore to learn useful representations, its inpainting task need masks that can erase semantically meaningful and difficult regions. Identifying meaningful regions or difficult examples without labeled data is extremely difficult. Similar ideas have recently been proposed by Gao *et al.* [10] and Wei *et al.* [44] to identify and use masks with different difficult levels for training, however, with a focus on handling arbitrary levels of corruption in semantic inpainting and weakly-supervised semantic segmentation

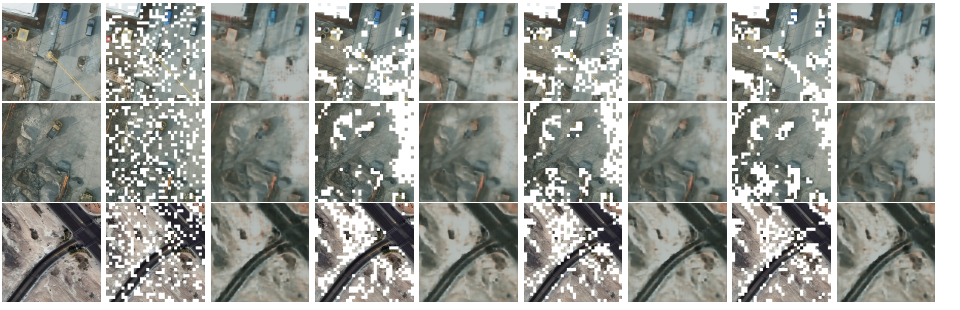


Figure 4: Coach model predicts an increasingly difficult masks for semantic inpainting. For each row, from left to right: Input image (512×512) for the coach network, masks predicted at iterations 0, 1, 6, and 8 with corresponding inpainting output. Note that at iteration 0 the coach predicts random masks.

with a pre-trained model, respectively. In contrast to using random mask from pre-defined distributions in [10], the coach network learns to score the regions based on difficulty in its inpainting. The coach is trained with loss adversarial to the reconstruction loss. In this way, the coach learns to create increasingly difficult examples for the inpainting network.

We propose *coach network* that learns a *semantically meaningful* mask M for the given image x (see Figure 3). The coach model C learns to assign meaningful score to the regions in image x by **maximizing** the reconstruction loss.

$$L_{coach}(x) = 1 - \mathcal{L}_{rec}(x \odot M) = 1 - \mathcal{L}_{rec}(x \odot C(x)) \tag{5}$$

However, applying this loss naïvely would result in the masks having 0 value at all regions because then no context information is present for the inpainting model and maximum reconstruction loss is achieved. Therefore, we apply constraints on outputs of the coach model to ensure a constant fraction of the images is always available as context for inpainting.

$$\hat{B}(x) = B(x) - \texttt{SORT}(B(x))^{k|B(x)|} \tag{6}$$

$$M = C(x) = \sigma(\alpha \hat{B}(x)) \tag{7}$$

The backbone network B of the coach model C has the same architecture as the encoder network of inpainting model. This gives the coach approximately similar representation power as the encoder network. $\texttt{SORT}(B(x))$ represents the sorting operation in descending order over all values in the activation map. $|B(x)|$ denotes the spatial size of activation map, k represents the k^{th} element in the sorted list of scores and controls the fraction of image to be erased. $\hat{B}(x)$ gives the relative difficulty score for each region with respect to the k^{th} element. The regions with score lesser than the k^{th} element are erased from the image while the other regions are kept intact. For example, $k = 0.75$ would erase $\frac{1}{4}$ area of the image. We scale the scores to the range [0, 1] using point-wise sigmoid function $\sigma(\alpha \mathbf{x})$, where α is a scalar that controls the steepness of σ. High α value results in discrete masks value $\{0, 1\}$ (for inpainting mask), whereas low α results in continuous mask values [0,1] (for training coach model). We use $\alpha = 1$ while training the coach network, and step-function ($\alpha \to \infty$) while training the inpainting network.

3.3 Training

We train coach and inpainting networks in an alternate fashion creating a competition between the models. The coach model learns to create increasingly difficult examples for the

inpainting model while the inpainting model learns superior feature with more difficult examples. The overall training objective (ignoring $\mathcal{L}_{con}$ for simplicity) is given by

$$L(x) = \min_{\theta_F} \max_{\theta_C} ||x - F(x \odot C(x, \theta_C), \theta_F)||_2^2 \tag{8}$$

where θ_F and θ_C are the parameters of the inpainting network and coach network, respectively. To introduce diversity and stochasticity in mask prediction, we inject noise sampled from a standard normal distribution to the coach's penultimate activation maps with the help of reparameterization [22]. In the first iteration of training inpainting model, we fill the mask (output of the coach network) with values drawn from a uniform distribution, $B(x)^{j,k} \sim U[0,1]$. We use this random mask as a starting point, instead of random patch mask as used in [36], to keep the nature of corruption same across iterations as semantic inpainting tends to overfit to the type of corruption it has been trained for [10]. Figure 4 shows few examples of meaningful and increasingly difficult masks predicted by the coach network.

4 Experiments and Results

4.1 Implementation Details

Semantic Inpainting We use input size of 128×128, batch size of 128 and employ random crops, mirroring, resizing, horizontal flip, and rotations for data augmentation. We empirically set $\mathbf{w}_{rec} = 0.99$ and $\mathbf{w}_{con} = 0.01$ in all experiments and find it to be a good balance between inpainting and learned feature quality. We use MSE loss clipped at 2 and observe that it allows the network to converge faster, predict pixel intensities far from the mean of the distribution. We use SGD optimizer with 0.9 momentum and 0.0005 weight decay to train the inpainting network for 100 epochs and step LR starting at 0.1 with step size 0.1.

Coach Networks Inputs, data augmentation, and batch size for this network is kept same as inpainting network. We remove the `maxpool` layer from ResNet-18 to predict the mask at a resolution of 8×8 and then apply $16\times$ nearest neighbor upsampling to scale the mask to 128×128. We erase 25% of the patches or 16 patches ($k = 0.75$) based on the predicted difficulty score. For Context Encoders [36], we remove 16 random patches of size 16×16 from the image. We train the coach network with Adam optimizer [21] at a fixed learning rate of 10^{-5} for 30 epochs at a time. This is followed by training of inpainting network for 30 epochs at a fixed learning rate of 10^{-5}. We repeat this procedure for 10 iterations.

Semantic Segmentation We adapt the inpainting network for semantic segmentation by removing the pixel-wise regressors. For the variant of inpainting network with bottleneck, following Long *et al.* [26], we apply a pixel-wise classifier at 3 scales: $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$. For the variant of inpainting network without the bottleneck, we apply a pixel-wise classifier at all 5 scales. We train all segmentation networks for 100 epochs and step learning rate starting at 0.001 with step size 0.1. We use input size of 256×256, batch size of 64 and employ the same data augmentation used for the training inpainting network. We observe that training segmentation network using small amount of data with cross-entropy loss leads to variations in segmentation results between re-runs. We train the segmentation network with soft-IOU loss [29] which leads to more stable and reproducible results. Readers are suggested to refer to the supplementary materials for additional implementation details.

Dataset	Resolution	Ground Resolution	Train	Validation	Crop Size	Stride	Task
Potsdam [19]	6000×6000	5 cm	20	4	600×600	200×200	Scene Parsing
SpaceNet Road [38]	1300×1300	30 cm	2000	567	650×650	250×250	Road network
DeepGlobe Lands [6]	2448×2448	50 cm	803	171	612×612	228×228	Land cover
DeepGlobe Roads [6]	1024×1024	50 cm	6226	1243	512×512	256×256	Road network
fMoW [5]	variable	50 cm	100000	2000	512×512	non-overlapping	Pre-training

Table 1: Statistics and other details for the datasets used in our experiments. We use non-overlapping crops for validation images for all datasets.

4.2 Datasets

We validate our ideas by performing experiments on four disparate datasets of overhead imageries with variations in the task, dataset size, and image ground resolutions. Note that, we use only 3 band RGB images in our experiments. The statistics of these datasets are given in Table 1. Readers are requested to see the the supplementary materials for examples and additional details of these datasets.

Potsdam [19] This dataset is used for scene parsing of the Potsdam city with 6 classes: `impervious surface`, `building`, `tree`, `low vegetation`, `car`, and `BG`.

SpaceNet Road [38] This dataset is used for road network estimation. The annotations are in the form of line-strings of roads. We obtain the binary masks by dilating the foreground to 40 pixels, resulting in road masks of roughly 12 meters width.

DeepGlobe Lands [6] This dataset is used for land cover estimation with 7 classes: `urban`, `agriculture`, `range`, `forest`, `water`, `barren`, and `unknown` (ignore class).

DeepGlobe Roads [6] This dataset is used for road network estimation. Pixel-level annotations are provided for `road` and `background` classes.

Functional Map of the World [5] We use only the images from the `train` split of this dataset to study the feature quality learned with respect to the number of unlabeled examples.

4.3 Results

We initialize all parameters with the technique proposed by He *et al.* [12]. We use mean Intersection-Over-Union (mIOU) as the metric at different amount of labeled and unlabeled data used for training. We do **not** apply weights to loss with respect to class population in any experiment and found that pre-training helps in alleviating the effect of class imbalance which is a prominent issue in overhead imagery tasks. Table 2 shows the performance of the baselines and our method while using all training images for the self-supervised pipeline and 10% labeled images of respective training set for fine-tuning the segmentation network. Figure 5 shows qualitative segmentation results of prediction from a model trained from scratch and a model pre-trained with our method.

Self-supervised Baselines We compare our results with three competitive self-supervised feature learning techniques: (a) Context Prediction [8], (b) Context Encoders [36], and (c) Splitbrain Autoencoders [45]. To evaluate their relative performance, we keep the AlexNet [23] architecture same for all the methods (see supplementary materials). Context Prediction and Context Encoders both tasks try to learn the structural information in the image, however, Context Encoders perform better in all cases, confirming semantic inpainting task being relatively closer to semantic segmentation task. Splitbrain AE outperforms Context Prediction and Context Encoders, confirming the findings of [45].

Method	Encoder	Bottleneck	Decoder	Results			
				Potsdam	SpaceNet	DG Roads	DG Lands
Context Prediction [8]	AlexNet	✗	✗	0.273	0.593	0.478	0.257
Context Encoders [36]		✓	✗	0.298	0.610	0.478	0.339
Splitbrain AE [45]		✓	✗	0.265	0.641	0.482	0.411
ImageNet	ResNet-18	✗	✗	0.493	0.701	0.669	**0.575**
Scratch	ResNet-18	✗	✗	0.414	0.657	0.643	0.495
Scratch		✗	✓	0.418	0.661	0.607	0.507
Autoencoder		✓	✓	0.502	0.748	0.749	0.515
Autoencoder		✗	✓	0.499	0.742	0.742	0.499
Context Encoders (Ours)	ResNet-18	✓	✗	0.540	0.730	0.478	0.501
		✗	✓	0.562	0.762	0.759	0.503
Coach Mask (Ours)	ResNet-18	✗	✓	**0.568**	**0.770**	**0.768**	0.529

Table 2: Semantic segmentation results (mIOU) while using full training set for the self-supervised pipeline and 10% of labeled images of respective datasets for training the segmentation network.

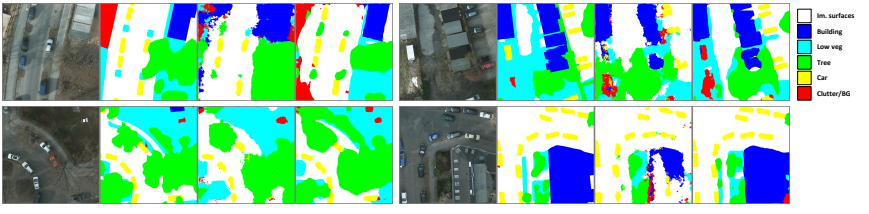
Architectural improvement ResNet-18 pre-trained on ImageNet performs better than training from scratch (see Table 2). This can be explained with the fact that the weights of earlier layers are generic and rarely change across domains. However, pre-training on ImageNet performs worse than simple autoencoder pre-training suggesting the large gaps between ground and overhead imageries. Table 2 also shows that having no bottleneck and re-using the pre-trained decoder network along with the encoder significantly improves the results, specially for road network extraction.

Interestingly, for DG Lands, pre-training on ImageNet performs better than unsupervised and self-supervised pre-training. We hypothesize that this is because image reconstruction and inpainting of the images used for land cover classification is inherently equivalent to texture completion leading to inferior self-supervision. Superior results with Splitbrain Autoencoders [45] cross channel prediction, among the baseline methods, further confirms that color and texture plays a major role in this task.

Learned masks Adversarial inpainting with increasingly difficult masks outperforms the baselines in all the tasks simultaneously (see Table 2 and 3). These improvements against the strong baselines, although seems small, is significant primarily because performance gain over sophisticated data augmentation is difficult. Note that, the domain gap in inputs between inpainting and segmentation is similar in cases of random and adversarial masks since an equal amount of region is erased from the input image.

Dataset	Method	(a) Labeled				(b) Unlabeled					
		10%	25%	50%	100%	1K	2K	5K	10K	50K	100K
Potsdam	Scratch	0.418	0.502	0.544	0.582	NA	NA	NA	NA	NA	NA
	Context Encoders (Ours)	0.562	0.628	0.668	0.698	0.432	0.453	0.537	0.561	0.548	0.562
	Coach Mask (Ours)	**0.568**	**0.637**	**0.674**	**0.705**	0.446	0.469	0.541	0.563	0.566	0.565
SpaceNet	Scratch	0.661	0.720	0.748	0.766	NA	NA	NA	NA	NA	NA
	Context Encoders (Ours)	0.762	0.781	0.795	0.804	0.696	0.731	0.754	0.759	0.763	0.765
	Coach Mask (Ours)	**0.770**	**0.786**	**0.797**	**0.806**	0.709	0.731	0.757	0.770	0.774	0.774

Table 3: Segmentation performance (mIOU) using the proposed architecture (ResNet-18 encoder, no bottleneck, and decoder) with respect to the (a) fraction of labeled images used for fine-tuning and (b) number of unlabeled images used for self-supervised training with 10% labeled data for fine-tuning.



(a)

(b)

Figure 5: Qualitative semantic segmentation results for Potsdam (a) and SpaceNet Road (b), from left to right: input image, ground truth, prediction with model trained from scratch, and prediction with model pre-trained using our approach. 10% of labeled data is used for fine-tuning in all cases.

Number of labeled and unlabeled samples used As expected, there is a consistent improvement for all methods when the number of labeled images is increased (see Table 3). Our adversarial training strategy consistently outperforms others with respect to different amounts of labeled images used for fine-tuning.

Surprisingly, the performance of self-supervised pre-training remains mostly the same despite a significant increase in number of unlabeled images used for pre-training (see Table 3). This behavior is most likely due to domain gap between semantic inpainting and semantic segmentation task. Furthermore, the random mask based inpainting technique suffer more than our proposed technique when the number of unlabeled images used for pre-training is drastically reduced. These results also conclude that our adversarial training have similar advantages and disadvantages when compared to the Context Encoders [36], however, it performs better in all scenarios we tested.

5 Conclusions

In this work, we propose a unified semantic segmentation approach towards a variety of overhead imagery tasks. We employ self-supervised techniques for pre-training due to scarcity of labeled data and availability of a large number of unlabeled data. Experiments show that existing self-supervised techniques, focusing primarily on classification, are inefficient for semantic segmentation. Our proposed architectural changes (3.1) leads to significant improvements in various diverse overhead imagery tasks. This is largely due to the use of high capacity ResNet-18 [13] as the backbone network and the re-use of pre-trained decoder networks. Additional improvements over strong baselines are observed on training the inpainting network with an adversarial coach network (3.2). The coach model is able to predict an increasingly difficult mask leading to a more difficult self-supervised task. However, existing self-supervised techniques as well as our proposed method do not exploit the availability of large unlabeled images. These insights motivate us to further probe self-supervised learning techniques to unlock the true potential of self-supervision in our future works.

References

[1] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In *ACCV*, 2016.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *PAMI*, 2017.

[3] Favyen Bastani, Songtao He, Mohammad Alizadeh, Hari Balakrishnan, Samuel Madden, Sanjay Chawla, Sofiane Abbar, and David DeWitt. RoadTracer: Automatic Extraction of Road Networks from Aerial Images. In *CVPR*, 2018.

[4] Alexander Buslaev, Selim Seferbekov, Vladimir Iglovikov, and Alexey Shvets. Fully convolutional network for automatic road extraction from satellite imagery. In *CVPRW*, 2018.

[5] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CoRR abs/1711.07846*, 2017.

[6] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *CVPRW*, 2018.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[8] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[10] Ruohan Gao and Kristen Grauman. On-demand learning for deep image restoration. In *CVPR*, 2017.

[11] Lluis Gomez, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *CVPR*, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.

[15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS*, 2018.

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 2006.

[17] Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length and helmholtz free energy. In *NIPS*, 1994.

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[19] ISPR. Potsdam 2d semantic labeling contest. http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html.

[20] Michael Kampffmeyer, Arnt-Børre Salberg, and Robert Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *CVPRW*, 2016.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[24] Nataliia Kussul, Mykola Lavreniuk, Sergii Skakun, and Andrii Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *GRSL*, 2017.

[25] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[27] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *NIPS*, 2014.

[28] Gellert Mattyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Enhancing road maps by parsing aerial images around the world. In *ICCV*, 2015.

[29] Gellert Mattyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *ICCV*, 2017.

[30] Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *ECCV*, 2010.

[31] Volodymyr Mnih and Geoffrey E Hinton. Learning to label aerial images from noisy data. In *ICML*, 2012.

[32] Agata Justyna Mosinska, Pablo Marquez Neila, Mateusz Kozinski, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *CVPR*, 2018.

[33] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *CVPR*, 2015.

[34] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

[35] Sakrapee Paisitkriangkrai, Jamie Sherrah, Pranam Janney, and Anton van den Hengel. Semantic labeling of aerial and satellite imagery. *J-STARS*, 2016.

[36] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[37] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML*, 2011.

[38] SpaceNet Road. Spacenet on amazon web services (aws). https://spacenetchallenge.github.io/datasets/datasetHomePage.html, 2017.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[40] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.

[41] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 2010.

[42] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *CVPR*, 2015.

[43] Jan D Wegner, Javier A Montoya-Zegarra, and Konrad Schindler. A higher-order crf model for road network extraction. In *CVPR*, 2013.

[44] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, 2017.

[45] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017.

[46] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *GRS Magazine*, 2017.