Land Cover Classification of Multi-Source Airborne Data using Conventional and Deep-Learning-Based Unsupervised Domain Adaptation

Edwin Deisling, Raphael Zipperer, Benedikt Kottler, Kevin Qiu, Melanie Böge, Dimitri Bulatov

Fraunhofer IOSB, 76725 Ettlingen, Germany - (name.surname)@iosb.fraunhofer.de

Keywords: 3D, airborne, generative, semantic segmentation, style transfer, training data.

Abstract

For an increasing number of applications, land cover maps can be generated from remote sensing imagery using conventional and deep-learning-based semantic segmentation models. Relying on a large pool of training data, the networks struggle with the spatial-temporal-spectral heterogeneity in the complex and diverse remote sensing imageries, leading to a significant number of errors in the model predictions. This paper presents a workflow comprising domain adaptation and classification. In particular, we analyze two domain adaptation techniques: First, a conventional histogram-matching method, which has turned out to be a surprisingly fast and reliable tool in a previous study, and second, a CycleGAN, which we applied both in its standard form and with the perceptual loss, thereby penalizing style inconsistencies on deeper layers. By applying the workflow to three remote sensing datasets and six directions of domain adaptation, we show that there is “no free lunch” in the sense that all domain adaptation methods have their advantages. Depending on the dataset, classification method, and especially on the availability of 3D data, the performance gap can be reduced to up to 1.5% of the mean F1 score, demonstrating the soundness of the proposed method.

1. Introduction and previous work

1.1 Motivation

Remote sensing (RS) deals with obtaining information about the Earth’s surface using airborne sensors. High-resolution images with multispectral channels can be captured, providing greater spectral and spatial detail as well as an increasingly rich information content. These techniques provide significant value for land cover classification, a term referring to the categorization of landscape elements to a class, such as buildings, vehicles, vegetation, and other types of objects. Land cover classification plays a crucial role in various domains, such as environmental monitoring to track changes in the ecosystem, city planning for sustainable urban development, and defense applications for strategic terrain analysis. However, not only because of the variety of sensors, varying image acquisition conditions, noise, and shadows, but also due to intra-class variations and inter-class similarities, land cover classification remains a challenging task (Wittich and Rottensteiner, 2021; Bulatov et al., 2019). Three-dimensional (3D) data, if additionally available, can provide a certain relief, because distances and elevations are usually given in absolute, metric values, but also here, there are two shortcomings. First, an effort to acquire them is considerable, because it involves either additional, costly sensors (ALS equipment) or computationally-intensive image processing methods (Photogrammetry, Structure from Motion). Second, for advanced classifiers, the nature of points matters so that the models are not transferable from one dataset to another. This already holds for conventional classifiers, such as Random Forest (RF, Breiman (2001), where the user can provide sophisticated features based on local and global neighborhoods. RF, known to fit hyperrectangles in the feature space that are possibly homogeneous regarding the target classes, needs many parameters and strongly depends on the similarity between (training and test set of) input features distributions, even if we deal with the same classes. Hence, models often struggle when applied to new data that differs from that used for training. In deep-learning (DL)-based classifiers, like Convolutional Neural Networks, very deep layers are supposed to learn how to interpret high-level context. Thus, the adaptability of such networks to the new data is known to be somewhat better than that of RF; however, it comes at the cost of significantly more complex models. For example, the ResNet (He et al., 2016) pipeline, which serves as the backbone for DeeplabV3+ (Chen et al., 2017), one of the best tools for semantic segmentation, has a considerably higher number of parameters, requiring plenty of labeled data.

1.2 Problem formulation

The formulation of the problem we are going to deal with for the rest of the paper is simple and typical for remote sensing: a model M is trained on a source dataset, where labeled data is available, and remains fixed. Then, M must be applied to a different target dataset, where the data is unlabeled. The source and target datasets will be denoted as $\mathcal{S}$ and $\mathcal{T}$, respectively, and in both cases, we refer to image features.

There exist many sensible ways to address this problem. We could, for example, label the data interactively. However, though excellent tools exist, such as (Russell et al., 2008), this proves to be highly time-consuming and challenging, as the datasets are extensive. Rasterizing available GIS data, such as building outlines or road centerlines (Bulatov et al., 2019), only works if the semantic classes coincide with the GIS data labels. Alternatively, we could approach unsupervised learning, where the goal is to learn the underlying structure of a dataset without labeled examples (Sarker, 2021), or self-supervised learning, where foundational models like DINO (Caron et al., 2021) are increasingly being use to learn better, more context-rich features by distillation, so that they may be used for meaningful clustering. Nevertheless, without knowledge of the classes, semantic information cannot be obtained from them. In all cases, the previous training effort for the existing model M would be wasted. Therefore, the so-called transfer learning aiming to improve the prediction on $\mathcal{T}$ by exploiting the knowledge from the input dataset $\mathcal{S}$.

1.3 Related work

According to the three frequently cited surveys of Tuia et al. (2016); Wang and Deng (2018); Xu et al. (2022), we could identify four to five main categories of methods for DA because hybrid methods play an important role since the boundaries between the main categories are often fuzzy.

The first two categories address the selection of either features (1) or data points (2) that exhibit a certain robustness regarding their change from the source to the target domain (Bruzzone and Persello, 2009; Cai et al., 2024; Han et al., 2024). The data-invariant samples can also be synthetic (Cai et al., 2023). Thirdly, one could adjust the classifier by fine-tuning predictions using a few labeled samples of $\mathcal{T}$, see (Bruzzone and Prieto, 2002; Huang et al., 2024). The disadvantage of this group of methods is that adding new data references is costly. Interestingly, Wang et al. (2006) manipulates the model in a way to achieve the lowest entropy with the given target samples; hence, no source samples are needed.

Finally, the most popular category of methods is called domain adaptation. The goal is to adapt the feature set of $\mathcal{T}$ to that of $\mathcal{S}$ so that the model can be successfully applied to the transformed feature vectors $(G_S(t), t \in T)$[1]. Among conventional unsupervised domain adaptation methods, histogram matching (Böge et al., 2025), data alignment with (kernel) Principal Component Analysis, (Nielsen and Canty, 2009)), or Canonical Correlation Analysis (Volpi et al., 2015) can be mentioned. Transformation into a new, latent space is often implicitly combined with denoising, opening the way to sparse representations and low-rank reconstructions, which are meant to avoid the influence of outliers and noise in the source domain samples. Manifold alignment methods apply the maximum mean discrepancy method to the non-linear projections of the feature vectors to adjust the distributions of source and target domains. Three types of manifold learning, according to Liu et al. (2021), are locally linear embeddings (where k-nearest neighbors are used to reconstruct low-dimensional embedding), Laplacian Eigenmaps (penalizing the distances between high distances of transformed points that are close to each other in the original space), and local tangent space alignment, extracting tangent space in latent spaces and aligning these tangent spaces of the source and target domains. All this is done to maintain the characteristics of the original data structure in the transformed domain, for which Liu et al. (2021) has proposed a manifold regularization framework.

Among deep-learning-based approaches, Ganin and Lempitsky (2015) implemented one of the first procedures. Their adversarial neural network architecture, consisting of a feature extractor, a label predictor, and a domain classifier, minimizes the so-called domain confusion loss within the training process. Generative methods presuppose (deep-learning-based) transformation
1) from $\mathcal{T}$ to $\mathcal{S}$, possibly, with finetuning the classifier (Tasar et al., 2020; Wittich and Rottensteiner, 2021; Ji et al., 2020) or
2) vice versa, much more seldom, because requires re-training M, see (Schenkel and Middelmann, 2020), however, without adversarial loss, or
3) both, requiring minimization of cyclicity losses (transform from one domain to another and back must match the first), in addition to adversarial losses transformation of $\mathcal{S}$ must resemble $\mathcal{T}$ (Schenkel et al., 2021; Soto et al., 2020).

[1] Mostly, we speak of unsupervised domain adaptation if there are no labels of the dataset $\mathcal{T}$, though sometimes it means that G is obtained without the knowledge of M.

The authors of Wittich and Rottensteiner (2021) balance the classification accuracy and generation quality by training the transferred samples with the same classifier as the original ones. Probably, this method can be also referred to as a hybrid method since the classifier is being optimized as well. There may be other penalty terms, such as those based on discrete Wavelet transformation (Zeng et al., 2024), allowing to integrate insigts in frequency domain, or “visiting loss”, intended to ensure that feature vectors of $\mathcal{S}$ are visited wherever possible, have also been proposed, see the original paper (Haeusser et al., 2017) and an application in RS domain (Chen et al., 2020). Finally, advanced concepts like attention and self-attention can be applied. For example, a multi-level attention mechanism, which includes a feature-level attention generated by shallow features and an entropy-level attention produced by a deep discriminative feature, was presented in (Zheng et al., 2020). In this regard, the fastest progress is achieved by self-learning algorithms, often based on visual transformer outputs that have been pretrained on massive data sets.

1.4 Contributions

Motivated by the research of Böge et al. (2025), we are interested in Domain Adaptation for remote sensing images with optical and DSM-based data. In this research, while carried out for two datasets and, hence, merely two directions of DA, Histogram Matching (HM) has turned out to be a surprisingly fast and robust tool for performing this task. Image-generating methods, a product of Generative Adversarial Networks (GANs) (**?**), play an important role in our society, for example, in the appearance of deepfakes in certain media, but also in RS, for the tasks of inpainting and style transfer. In this paper, we will provide a comparison of the conventional and easily implementable HM method with a more sophisticated generative-adversarial network CycleGAN. This method will be extended by perception loss (Johnson et al., 2016), allowing to take into account high-level features.

We generalize our experiments for a third dataset, a well-known ISPRS Potsdam benchmark that was prepared in the same way as both previously considered datasets. The number of directions for DA is now six, which is sufficiently generalized. Similar to Böge et al. (2025), we wish to investigate the effect of additional features, such as the relative elevation and near infrared channel, as well as a conventional (RF) and a deep-learning-based classifier (DeepLab v3+), on the classification accuracy.

2. Methodology

In the following two subsections, we briefly cover the methods for domain adaptation and classification. Since these are established methods, we restrict ourselves to concise descriptions and will mostly mention the most important algorithm parameters. The only exception is the CycleGAN method, where the implementation of perceptual loss requires more detailed insight of the standard network.

2.1 Domain adaptation

2.1.1 Histogram matching An image feature is a one-dimensional variable. We consider its probability distribution and, therefrom, the cumulative distribution function CDF. A discretized histogram of a CDF is a monotonously increasing function mapping the feature space to the interval $[0; 1]$, and,

using a certain relaxation technique, we obtain a strictly monotonously increasing function. Since every strictly monotonous function is invertible, we define the HM as a transformation G

$$G(s_j) = \mathrm{CDF}_T^{-1}\mathrm{CDF}_S(s_j), \tag{1}$$

applied to every bin j of every feature s of the dataset $\mathcal{S}$. The algorithm parameters are the number of bins and the kind of relaxation. The `uniform` method uses a direct and exact intensity mapping based on the histogram, as described previously, while `polynomial` smooths the mapping through polynomial approximation by using cubic Hermite polynomials (Fritsch and Carlson, 1980), which proves to be suitable for images with large differences in brightness. To minimize the effects of overfitting, we consider 256 bins for the images of dimension four since we adapt the color and the near infrared channels.

2.1.2 CycleGAN A GAN is an architecture consisting of two neural networks: a Generator (G) and a Discriminator (D). In each training iteration, the Discriminator learns to distinguish between generated and ground truth images while the Generator is trained to maximize the Discriminator loss. CycleGAN (Zhu et al., 2017) expands on this architecture in the setting of style transfer between two feature sets $\mathcal{S}$ and $\mathcal{T}$, with one GAN for each respective direction. In other words, we train two generators $G_\mathcal{S}, G_\mathcal{T}$ and two discriminators $D_\mathcal{S}, D_\mathcal{T}$.

To couple them, CycleGAN introduces an additional cycle consistency loss, measuring divergence between a ground truth image (e.g. $\mathcal{S}$) and the result of subsequently applying both Generators $\tilde{\mathcal{S}} = G_\mathcal{S}(G_\mathcal{T}(\mathcal{S}))$, which should ideally result in the identity function. Penalizing differences between $\mathcal{S}$ and $\tilde{\mathcal{S}}$ over pixels is referred to as L_1 loss.

Besides this commonly used loss, symmetrically averaged over $\mathcal{S}$ and $\mathcal{T}$, we introduce an additional perceptual loss term into the cycle consistency loss. Perceptual loss L_perc from Johnson et al. (2016) minimizes the L_1 distance between feature representations in a pre-trained CNN of generated and ground truth image, making use of high level features and providing some invariance towards slight shifts. For feature extraction in perceptual loss, we use the VGG19 network. Hereby, if the input is not a three-channel image, as in our case (RGB+NIR, see Section 3.1), we can treat the NIR channel as a gray image since the difference to the visual spectrum is not significant and VGG19 accepts gray images as input, too. Finally, CycleGAN also utilizes identity loss L_id penalizing differences between $\mathcal{S}$ and $G_\mathcal{S}(\mathcal{S})$. According to Liu (2022), this loss aids in finding features in the domain $\mathcal{S}$ and preserving them during inference. The complete loss function is symmetric and is depicted in the following equation:

$$\begin{aligned} &L = L_\mathcal{S} + L_\mathcal{T}, \text{ with } L_\mathcal{S} = L_\text{adv}(G_\mathcal{S}, D_\mathcal{S}) + \\ &(\lambda_1 L_1 + \lambda_\text{perc} L_\text{perc}) \circ (\tilde{\mathcal{S}}, \mathcal{S}) + \lambda_\text{id} L_\text{id}(G_\mathcal{S}, \mathcal{S}) \end{aligned} \tag{2}$$

and an analogical $L_\mathcal{T}$. In (2), the first summand of $L_\mathcal{S}$ represents the typical for GANs adversarial loss, while the second summand penalizes the deviations between low-level and high-level features of $\mathcal{S}$ and those of its twofold reconstruction. The individual contributions are weighted by the factors $\lambda_1 = 0.95$, $\lambda_\text{perc} = \lambda_\text{id} = 0.05$ while for the vanilla cycleGAN, $\lambda_\text{perc} = 0$. We adopted the PatchGAN of Isola et al. (2017). Its generator is an encoder-decoder structure with a U-Net 256 as the backbone. There are eight layers in total, and the number of features on the deepest, most context-rich layer equals 512. The discriminator consists of five layers with convolutional, batch-normalization, and non-linear units. It is patch-based (giving the method its name); hence, there is no single “real / fake” output scalar for the whole image. Instead, the discriminator maps the input (real or generated) to a 2D grid of scores. Thus, each score judges a small patch (receptive field) of the input image. The individual scores are aggregated using, e.g., MSE loss. Compared to the VGG19 net used to compute perceptual loss (and justifying its application), the discriminator is rather shallow; hence, it enforces local realism (textures, local consistency) without penalizing global structural errors, such as object placement or large deformations, too much.

Parameter	Moabit ↔ Munich	Munich ↔ Potsdam	Potsdam ↔ Moabit
LR	6e−6 / 6e−6	4e−5 / 4e−5	6e−6 / 2e−5
Epochs	110 / 80	90 / 55	90 / 55

Table 1. Learning rate (LR) and number of epochs by CycleGAN model trained in both directions for a pair of datasets. Left and right from the slash networks with L1 loss only resp. those with perceptual loss are referred to.

The selection of the optimal learning rate and training epochs, indicated in Table 1 for completeness, was based on visual assessment of the results to ensure effective model performance and avoid overfitting. For each epoch, we employ interleaving for training first both generators and then the discriminators and rely every time on the Adam optimizer.

2.2 Land cover classification

2.2.1 Random Forest Random Forest (Breiman, 2001) is one of the most popular conventional classifiers when it comes to processing multi-modal data. A decision tree, with branches corresponding to splits across randomly chosen features, is supposed to fit hyper-rectangles in the labeled data set in order to obtain a possibly pure subdivision. After training of several such trees, they are applied to the non-labeled samples to retrieve probabilities of single classes. Except for the number of trees (which are supposed to reduce the bias), the minimum leaf size, corresponding to the maximum acceptable rate of impurity, is the key parameter. For the sake of speed, we decomposed the image into superpixels of approx. 0.25 m^2 each using the SLIC algorithm (Achanta et al., 2010). Then we trained and evaluated RF on superpixels using 20 decision trees while the minimum leaf size parameter was set to four. Two configurations are used: pure 2D (RGBIR channels) and a 2D+3D configuration, meaning extension by NDSM and planarity features.

2.2.2 Multimodal DeepLab We use the DeepLabV3+ architecture (Chen et al., 2017) based on a ResNet101 backbone (He et al., 2016). To process multi-modal input, we extend this architecture with an additional branch as in (Qiu et al., 2022), which takes NIR, planarity (Gross and Thoennessen, 2006), and NDSM (Bulatov et al., 2012) information as additional modalities. Both branches are initialized using ImageNet weights and can be weighted by a factor α. When $\alpha = 1$, we use only the RGB branch. When α is set to 0.5, both branches are equally weighted, and $\alpha = 0$ means that only the second branch, consisting of NDVI (instead of NIR), planarity, and NDSM, is used. These three configurations of this network are trained on patches of size 512×512 pixels using cross-entropy loss. The learning rate of the ResNet encoder is lowered by a factor of 10 compared to the decoder part of DeepLab. Unlike Böge et al. (2025), who used the ISPRS benchmark Potsdam (Rottensteiner et al., 2014) dataset for pre-training, we rely on

$\mathcal{S} \leftarrow \mathcal{T}$	Munich	Moabit ← Munich				Potsdam ← Munich			
DA mode	-	None	HM	cGAN	cGANp	None	HM	cGAN	cGANp
RF 2D	69.79 / 46.07	39.62 / 17.72	46.60 / 28.33	47.32 / 29.03	**49.05 / 30.06**	50.79 / 34.33	51.78 / 34.54	55.30 / 37.32	**56.19 / 38.91**
RF all	85.98 / 65.63	53.68 / 33.22	75.67 / 56.20	79.45 / 59.76	**79.76 / 60.15**	75.56 / 54.91	77.48 / 52.01	80.56 / 57.11	**80.77 / 58.86**
DL RGB	82.42 / 67.87	52.67 / 35.91	**63.01 / 56.04**	59.23 / 48.53	60.29 / 49.66	65.58 / 47.63	**66.47** / 45.05	66.02 / 48.50	65.67 / **48.93**
DL ~~RGB~~	83.09 / 65.69	78.86 / 55.50	81.85 / 59.62	**83.16 / 64.12**	82.68 / 63.00	74.03 / 51.54	77.04 / 50.76	79.89 / 55.75	**80.21 / 56.93**
DL all	84.70 / 69.57	67.72 / 54.66	82.59 / 64.39	**83.34 / 67.12**	83.24 / 65.78	74.81 / 56.90	79.31 / 55.76	**80.24** / 59.09	80.00 / **59.87**

$\mathcal{S} \leftarrow \mathcal{T}$	Munich ← Moabit				Moabit	Potsdam ← Moabit			
DA mode	None	HM	cGAN	cGANp	-	None	HM	cGAN	cGANp
RF 2D	45.46 / 29.86	**52.11 / 35.87**	50.17 / 32.62	49.54 / 32.92	69.50 / 53.75	31.90 / 20.50	43.70 / 27.95	50.68 / 34.99	**51.17 / 36.82**
RF all	79.70 / 57.02	**81.87 / 58.18**	81.38 / 57.65	81.81 / 57.68	84.50 / 72.99	54.09 / 37.41	78.73 / 54.07	77.13 / 54.56	**78.95 / 57.66**
DL RGB	33.71 / 25.64	**70.80 / 58.82**	62.49 / 50.65	68.94 / 55.70	88.89 / 73.59	27.12 / 20.43	67.53 / **58.27**	58.02 / 42.56	**68.15** / 52.15
DL ~~RGB~~	87.45 / 62.00	**87.59 / 62.04**	84.44 / 54.43	84.08 / 53.90	89.71 / 69.58	87.36 / **58.37**	**87.67** / 57.08	79.99 / 47.81	82.96 / 51.26
DL all	88.22 / 67.23	**88.49 / 68.61**	84.70 / 63.41	86.85 / 64.86	90.94 / 75.55	61.57 / 51.22	**86.36 / 68.28**	84.17 / 60.84	85.53 / 64.57

$\mathcal{S} \leftarrow \mathcal{T}$	Munich ← Potsdam				Moabit ← Potsdam				Podstdam
DA mode	None	HM	cGAN	cGANp	None	HM	cGAN	cGANp	-
RF 2D	51.16 / 32.75	48.10 / 30.99	**59.38 / 38.78**	57.57 / 37.68	34.30 / 11.56	42.11 / 30.10	45.93 / 28.32	**49.45 / 32.67**	69.74 / 52.11
RF all	72.46 / 52.50	68.10 / 48.28	71.82 / 51.00	**71.89 / 51.27**	43.07 / 22.57	64.52 / 46.86	66.87 / 48.41	**68.14 / 49.46**	82.39 / 65.04
DL RGB	68.02 / 52.11	56.64 / 46.72	**72.01 / 58.25**	69.22 / 57.27	40.59 / 28.55	52.61 / 45.83	52.48 / 44.88	**58.87 / 50.06**	84.08 / 71.07
DL ~~RGB~~	64.26 / 47.10	**72.41 / 52.00**	68.29 / 49.94	67.57 / 49.33	68.49 / 49.35	71.86 / **52.98**	72.32 / 52.00	**73.78** / 52.91	84.06 / 63.93
DL all	68.26 / 54.60	67.01 / 52.58	**73.44 / 57.25**	71.58 / 55.72	59.03 / 45.17	68.77 / 51.18	77.36 / **58.54**	**77.71** / 57.69	86.44 / 71.61

Table 2. Results of domain adaptation. The target dataset ($\mathcal{T}$) needs to be adapted to the source dataset ($\mathcal{S}$) since the latter has a classification model. In the rows, different pairs of datasets with fixed $\mathcal{T}$ and different DA methods. In the rows, different classification models: RF 2D and RF all signify relying on radiometric image channels (R, G, B, IR), respectively, additionally with NDSM and planarity; DL RBG, DL ~~RGB~~, and DL all denote choices of α to be 1, 0, and 0.5, respectively. OA and mF1 are recorded to the left and to the right of the slash, respectively. The best result for a given classification model and the pair of datasets is marked in bold, while the grey cells show the most impactful reduction of PG.

the ImageNet dataset (Russakovsky et al., 2015), because Potsdam was one of the datasets considered in this work. In the follow, we use the abbreviation DL for multi-modal DeepLabV3, as counterpart to RF.

3. Results

3.1 Datasets and experimental setup

Munich and Moabit, already discussed in Böge et al. (2025), were extended by the Potsdam dataset, prepared in the same way as both aforementioned ones: We have an orthophoto with four channels (RGB + NIR, 2D data) and NDSM (3D data). We compute NDVI and planarity from the orthophoto and the NDSM, respectively. Besides, 40 512×512 pixel patches were labeled Potsdam is an ISPRS benchmark dataset and has a particular problem: building margins exhibit strong artifacts in both 2D and 3D data. These artifacts were caused during orthophoto creation and were often labeled as the cluster class. Also in the areas of moving vehicles, the 3D data is mostly not accurate because this image sequence was reconstructed using a photogrammetric procedure.

We performed domain adaptation from each dataset to each dataset and recorded in the columns of the Table 2 the measures of overall accuracy (OA) and average F1-Score (mF1) of the different DA methods: no DA at all, HM, and CycleGAN without and with perceptual loss (abbreviated as cGAN and cGANp, respectively). In the rows, we differentiate between our five classification models (RF and DL, both with and without 3D data), which remain fixed for each dataset. We compare the performance with that of each model trained on the target dataset and track this way the so-called performance gap (PG). Then, the best possible results are shown in subtables marked in bold. For this, we subdivided training and test patches in the ratio 3:1.

3.2 Quantitative results

Our first observations from Table 2 are that the relative decay in performance in the absence of the model for $\mathcal{T}$ is stronger in the mF1 than in the OA, indicating the necessity of training data in the target domain for achieving good performance for rare classes, and that both performance measures indicate similar trends in most cases, with differences that can be attributed to the noise. That said, one can focus on one of these metrics. Most importantly, we note that for all configurations of datasets, performance metrics, and classification models (with one exception, namely the model trained on the second branch of DL for Moabit with Potsdam data, probably also results from noise), it is better to perform one kind of DA than none.

Comparing the DA methods with each other, we observe that CycleGAN outperforms HM in most configurations, and that the performance further improves after considering additional losses (cGANp). However, all three DA methods have their advantages that can be observed across both datasets (gray cells in Table 2) and classifiers. Regarding classifiers, occasional triumphs of the HM method can be mostly observed for deep-learning-based classifiers. One possible interpretation is that the discriminator of our GANs is relatively flat, and hence, DL, which has a quite deep backbone, notices the inconsistencies of DA using cGAN in the deep layers to a larger extent than RF, which only analyzes raw channels. Using the perceptual loss, partly, though not always, helps to solve this problem. Positively, in most cases, cGANp is either the best or the second-best method with very insignificant distances to the winner, which confirms the additional value of the perceptual loss.

Further remarks on the classifiers are that the classification with the deep-learning-based classifier is, with one exception, better than with a conventional one, as can be expected (comparisons between the first vs third row and between the second vs last row of all tables). One exception is the overall accuracy metric

for the Munich-to-Potsdam configuration without any domain adaptation. Here, a better performance of RF with combined 2D and 3D data can be attributed to noise as well as the fact that the RF-based classification takes place on superpixels. It is also clear that adding features resulting from 3D data and an index (NDVI) improves the performance and reduces the differences between the single DA methods (comparisons between the first vs second row and between third vs last row of all tables). This is because these features are widely illumination-invariant and are not subject to DA. As a consequence, the results are more dramatic if the classifier relies on the image data only. However, in the fourth row of Table 2, for the configuration Moabit-to-Potsdam, significant differences in performance indicate the importance of DA even for the calculation of NDVI because all other inputs for the DL method remain fixed.

Regarding datasets, for the Moabit-to-Munich data transfer, HM performs best while for the Potsdam-to-Munich configuration, the winner is mostly the standard CycleGAN. The reason may be that to guarantee the fairness of the comparison, we widely to keep the parameters for GANs constant, since without reference labels for $\mathcal{T}$, it is not possible to finetune them, and for certain pairs of datasets, they may not be optimal. Furthermore, the Munich and Potsdam datasets, as qualitative results will show, look more similar to each other than to the Moabit dataset, and therefore, the results without the application of DA yield a reasonable performance. As soon as one of the datasets is Moabit, the performance varies strongly, sometimes by almost 40 percentage points of difference across the DA methods for the same classification model, for example, 20.43% without and 58.27% with DA for the DL RGB configuration classfication of the Moabit dataset with the Potsdam model, thus emphasizing the additional value of DA in the case of merely 2D data available. Munich is the least problematic dataset, probably, also because of the accurate DSM data provided by the SURE pipeline of Rothermel et al. (2012).

Overall, applying DA allows for reducing the performance gap in mF1 to less than two, seven, and 13 percentage points for the Munich, Moabit, and Potsdam datasets, respectively. This reduction can be achieved using the DL-based classifier with combined 2D and 3D features and different DA methods. The differences between datasets may be attributed to clutter regions (around buildings) in the Potsdam dataset and to the river in the Moabit dataset.

3.3 Qualitative results

Although we considered two classifiers for the quantitative analysis, we will now focus primarily on DL in the qualitative analysis. The reason for this is as follows: Similar to the DL classifier, the RF classifier can identify buildings, roads and trees. However, due to its underlying mechanism and the partition into superpixels, the classification of smaller objects like cars is less accurate and, in some cases, significantly more challenging. In contrast, the DL classifier does not exhibit this issue to such a large extent. Figures 1 and 2 show RGB patches of the three cities together with their adapted versions (we omitted the near-infrared channel). Depending on source and target dataset, one method works better than the other.

Starting with the Munich-to-Moabit configuration, we observe the good performance of HM. Buildings are classified more precisely, and roads can mostly be distinguished from buildings, whereas with CycleGAN adaptation, roads are sometimes incorrectly classified as buildings. Once elevation data are included, any method can resolve this discrepancy. By adapting the Munich features to the Potsdam features, the distinction between roads and buildings in 2D data improves slightly. However, the classification of buildings in general is less accurate and inconsistent, which subsequently can be improved by incorporating 3D information. The choice of method has little influence on the results, though the predictions of the CycleGAN-adapted images provide slightly more accurate results.

Considering the RGB channels of Moabit (second row), we notice, in particular, an improvement of the prediction results, regardless of which other city they are adjusted to. Without adapting the channels, grass or roads are often incorrectly recognized. Domain adaptation allows significantly more buildings to be detected and trees to be distinguished from grass. HM stands out in this regard, as it allows for more accurate predictions of grass and clutter.

The prediction results also improve on the Potsdam dataset as target $\mathcal{T}$ (bottom row). Depending on $\mathcal{S}$, the performance gains can be significant. Adapting using CycleGAN on the Munich RGB channels leads to a better class separability of trees and grass. Although the inclusion of 3D data improves the accuracy of road detection, it reduces the ability to distinguish between trees and grass. The situation is different in the case of Potsdam-to-Moabit. Without domain adaptation, the classifier mostly recognizes buildings rather than the other classes. All methods improve prediction, with cGANp standing out for its more reliable differentiation between trees and grass. The inclusion of 3D data leads to the same problems as in the Potsdam-to-Munich case, as HM adaptation frequently causes cars to be incorrectly recognized.

In Figure 2, we consider a challenging situation around the river in the dataset Moabit, where we track the result of domain adaptation with different source data (Moabit, Munich, Potsdam) and DA methods, while the classifier DL remains fixed with configuration $\alpha = 0.5$. It should be mentioned that the classifier does not recognize the river as a distinct class but instead assigns it to the clutter class, as it does not correspond to any of the other defined classes.

Starting with the first patch pair, we consider the original Moabit, where the prediction result is obtained by applying the model to the entire dataset. The second patch pair illustrates the Munich-to-Moabit case using HM as the DA (that is, inference with the Moabit model and $\alpha = 0.5$). Doing so yields the best result for cross-domain cases, as the classifications are largely correct. The third patch pair consists of patches adapted using cGAN and cGANp, along with their corresponding predictions. Clutter occurs more often in the cGAN adaptation, leading to river pixels being more frequently classified as clutter compared to cGANp.

The two right-most patches correspond to the case of inference with the Potsdam model, following the same arrangement as the previous two. Since the features are adapted to those of Potsdam, the classifier assigns the river to the road class, possibly because there was a road of approximately same color. As a result, the boat on the river can now be detected as a car. This has significant consequences for the classification of the images adapted using the CycleGAN methods: Adapting to the features of the Potsdam dataset results in trees being hallucinated on the road, leading to misclassified trees in the predictions. This example confirms that not only the adaptation method but also the source dataset significantly influences the prediction.

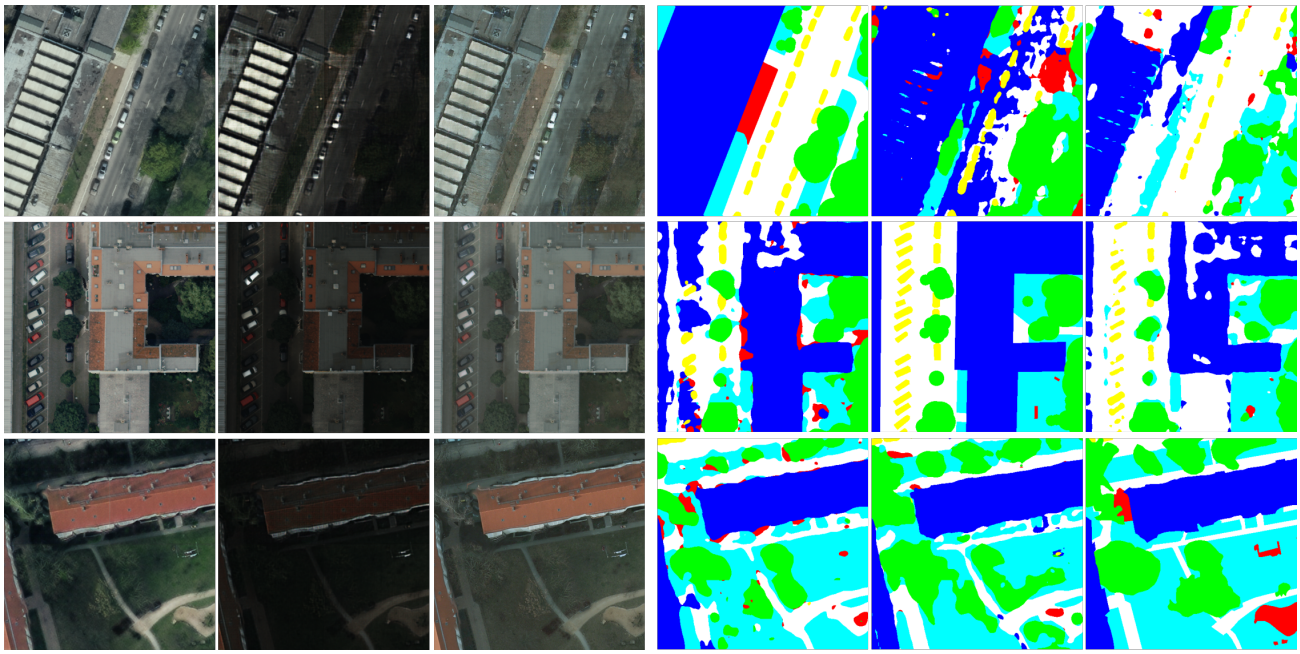

Figure 1. Left-hand side: Example patches of the three cities along with their adapted versions. The patches are arranged in the same way as in the Table 2. This means that the top left patch represents the original Munich, top middle and right: its adaptation to the Moabit and Potsdam models, respectively, etc. In these example patches, Munich (top row), Moabit (middle row), and Potsdam (bottom row) were adapted using cGAN, HM, and cGANp, respectively. Right-hand side: The corresponding prediction results for the patches on the left-hand side using DL and $\alpha = 1$, depicted with the standard colors of the Potsdam dataset (Rottensteiner et al., 2014).

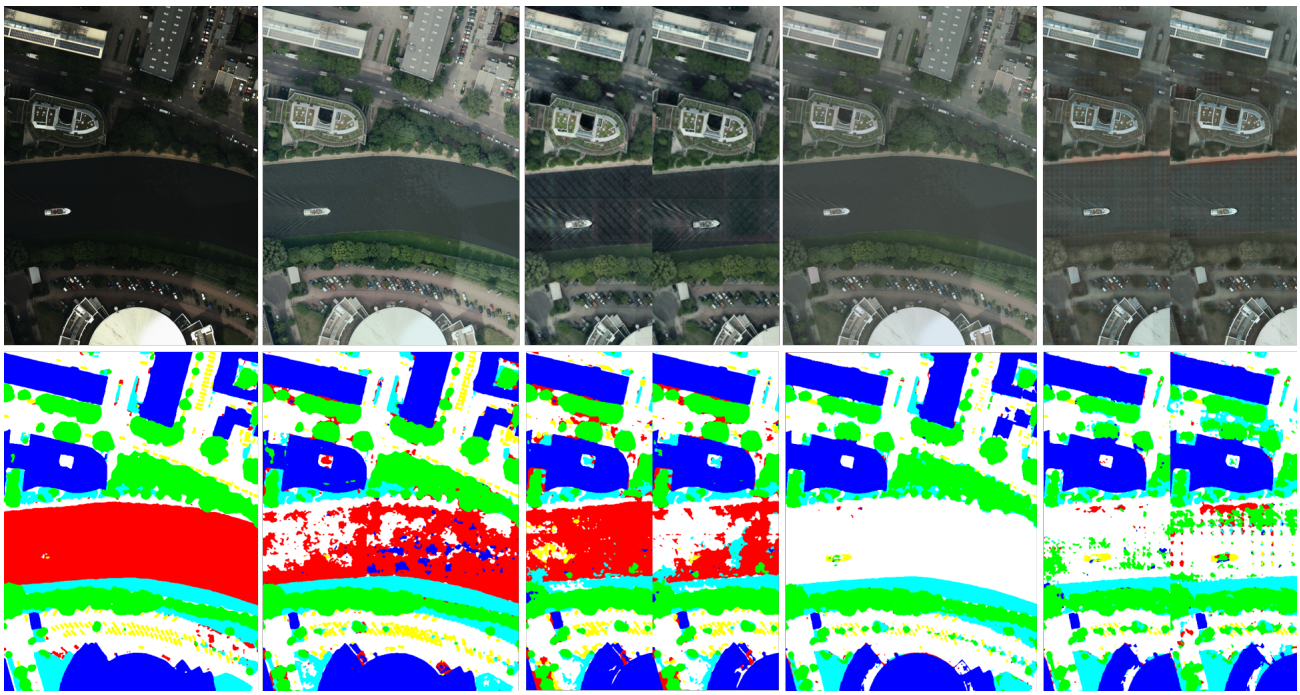

Figure 2. Original and adapted patches from the river that flows through Moabit. The first two adaptations are based on the features of Munich, and the last two according to the Potsdam features. The second and the fourth patch are adaptations using HM, while the third and the fifth patch are subdivided to contain results of both cGAN (always on the left) and cGANp (on the right). While clutter is detected at the river location in the patches adapted to Munich, the last two patches are instead classified as road in this region.

Note that apart from the river, all individual results, depicted in the bottom row, are similar and, to a large degree, good with respect to detection of large buildings, trees, and even cars. As noticed earlier, the adapted features of the Munich dataset are sometimes more anomalous than those of Potsdam, leading to slightly more clutter regions.

Summarizing, domain adaptation improves classification. Depending on which datasets are considered, one DA method may yield more accurate results than another. While HM performs a simple adjustment of color and brightness, CycleGAN enables a more semantically consistent adaptation. When cGANp outperforms cGAN, DL yields more consistent and semantically

meaningful predictions. Consequently, large buildings and long streets contain fewer incorrect labels, which ultimately results in more stable predictions.

4. Conclusions

Semantic segmentation of datasets without proper training data is one of the most important and best-explored problems in Machine Learning. We presented a workflow combining unsupervised domain adaptation and supervised land cover classification of optical and elevation-based remote sensing data. Currently, both conventional (Histogram Matching and Random Forest) and deep-learning-based tools (CycleGAN and DeeplabV3+) can be employed to perform land cover classification on datasets that lack training data, and, because of this modular structure, the workflow can be extended by new methods; for example, the CycleGAN method with perceptual loss. We applied this workflow to three datasets, resulting in six directions of DA, and found that all approaches have their place and advantages. In particular, the simple HM approach could mostly yield sound results. In other words, combining DL-based approaches could occasionally underperform, both because of possibly suboptimally chosen network parameters for some pairs of datasets or inconsistencies of DA in the deepest network layers. In future work, therefore, it would be important to combine the findings of both modules, for example, by supporting the DA with hints from the classifier on which features are most crucial for separating individual classes.

Moreover, in recent years, the amount and relevance of machine learning tasks and methods have increased with technical progress. However, mathematical proofs of lower bounds on classification error could not keep pace with this progress, so that theoretical exploration of the approaches in question could constitute an interesting topic for future research.

Acknowledgement

We would like to thank Dirk Frommholz, Dennis Dahlke, Magdalena Linkiewicz, Sebastian Pless, Karsten Stebner and Matthias Geßner from the Institute of Space Research of the German Aerospace Center (DLR) in Berlin/Germany for acquiring, pre-processing, and providing us the Moabit dataset. Labeling for the Munich and Moabit dataset patches was carried out by our student Lena Borchardt.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2010. SLIC superpixels. Technical report, Technical Report EPFL.

Böge, M., Bulatov, D., Deisling, E., Häufel, G., Qiu, K., 2025. Unsupervised domain adaptation for remote sensing data classification model transfer. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 17–24.

Breiman, L., 2001. Random Forests. *Machine Learning*, 45. https://doi.org/10.1023/A:1010933404324.

Bruzzone, L., Persello, C., 2009. A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability. *IEEE Transactions on Geoscience and Remote Sensing*, 47(9), 3180–3191.

Bruzzone, L., Prieto, D. F., 2002. A partially unsupervised cascade classifier for the analysis of multitemporal remote-sensing images. *Pattern Recognition Letters*, 23(9), 1063–1071.

Bulatov, D., Häufel, G., Lucks, L., Pohl, M., 2019. Land cover classification in combined elevation and optical images supported by OSM data, mixed-level features, and non-local optimization algorithms. *Photogrammetric Engineering & Remote Sensing, Volume 85, Number 3, pp. 179-195(17).*

Bulatov, D., Wernerus, P., Gross, H., 2012. On applications of sequential multi-view dense reconstruction from aerial images. *ICPRAM (2)*, 275–280.

Cai, Y., Shang, Y., Yin, J., 2024. Multidan: Unsupervised, multistage, multisource and multitarget domain adaptation for semantic segmentation of remote sensing images. *Proceedings of the 32nd ACM International Conference on Multimedia*, 1168–1177.

Cai, Y., Yang, Y., Shang, Y., Shen, Z., Yin, J., 2023. DASR-SNet: Multitask domain adaptation for super-resolution-aided semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848.

Chen, M., Ma, L., Wang, W., Du, Q., 2020. Augmented associative learning-based domain adaptation for classification of hyperspectral remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 6236–6248.

Fritsch, F. N., Carlson, R. E., 1980. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2), 238–246. http://www.jstor.org/stable/2156610.

Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, 1180–1189.

Gross, H., Thoennessen, U., 2006. Extraction of lines from laser point clouds. *Symposium of ISPRS Commission III: Photogrammetric Computer Vision PCV06. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 36number Part 3, 86–91.

Haeusser, P., Frerix, T., Mordvintsev, A., Cremers, D., 2017. Associative domain adaptation. *Proceedings of the IEEE International Conference on Computer Vision*, 2765–2773.

Han, J., Yang, W., Wang, Y., Chen, L., Luo, Z., 2024. Remote sensing teacher: Cross-domain detection transformer with learnable frequency-enhanced feature alignment in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, H., Li, B., Zhang, Y., Chen, T., Wang, B., 2024. Joint distribution adaptive-alignment for cross-domain segmentation of high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A., 2017. Image-to-image translation with conditional adversarial networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976.

Ji, S., Wang, D., Luo, M., 2020. Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5), 3816–3828.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. *European Conference on Computer Vision*, Springer, 694–711.

Liu, S., 2022. Study for Identity Losses in Image-to-Image Domain Translation with Cycle-Consistent Generative Adversarial Network. *Journal of Physics: Conference Series*, 2400, 012030.

Liu, X., Zhan, Z., Yuan, J., 2021. Domain adaptation algorithm based on manifold regularization. *IEEE International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, IEEE, 30–33.

Nielsen, A. A., Canty, M. J., 2009. Kernel principal component and maximum autocorrelation factor analyses for change detection. *Image and Signal Processing for Remote Sensing XV*, 7477, SPIE, 266–271.

Qiu, K., Budde, L. E., Bulatov, D., Iwaszczuk, D., 2022. Exploring fusion techniques in U-Net and DeepLab V3 architectures for multi-modal land cover classification. *Earth Resources and Environmental Remote Sensing/GIS Applications XIII*, 12268, SPIE, 190–200.

Rothermel, M., Wenzel, K., Fritsch, D., Haala, N. et al., 2012. Sure: Photogrammetric surface reconstruction from imagery. *proceedings LC3D workshop, Berlin*, 8number 2.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J., Breitkopf, U., Jung, J., 2014. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, 256-271.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al., 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.

Russell, B. C., Torralba, A., Murphy, K. P., Freeman, W. T., 2008. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1), 157–173.

Sarker, I. H., 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2.

Schenkel, F., Hinz, S., Middelmann, W., 2021. Style transfer-based domain adaptation for vegetation segmentation with optical imagery. *Applied Optics*, 60(22), F109–F117.

Schenkel, F., Middelmann, W., 2020. Domain adaptation for semantic segmentation of aerial imagery using cycle-consistent adversarial networks. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 1448–1451.

Soto, P., Costa, G., Feitosa, R., Happ, P., Ortega, M., Noa, J., Almeida, C., Heipke, C., 2020. Domain adaptation with CycleGAN for change detection in the Amazon forest. *ISPRS Archives; 43, B3*, 43, 1635–1643.

Tasar, O., Happy, S., Tarabalka, Y., Alliez, P., 2020. ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10), 7178–7193.

Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 41–57.

Volpi, M., Camps-Valls, G., Tuia, D., 2015. Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 107, 50–63.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T., 2006. Fully Test-time Adaptation by Entropy Minimization. ArXiv. 2020 doi: 10.48550. *arXiv*.

Wang, M., Deng, W., 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312, 135–153.

Wittich, D., Rottensteiner, F., 2021. Appearance based deep domain adaptation for the classification of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180, 82–102.

Xu, M., Wu, M., Chen, K., Zhang, C., Guo, J., 2022. The eyes of the gods: A survey of unsupervised domain adaptation methods based on remote sensing data. *Remote Sensing*, 14(17), 4380.

Zeng, J., Gu, Y., Qin, C., Jia, X., Deng, S., Xu, J., Tian, H., 2024. Unsupervised domain adaptation for remote sensing semantic segmentation with the 2D discrete wavelet transform. *Scientific Reports*, 14(1), 23552.

Zheng, J., Fu, H., Li, W., Wu, W., Zhao, Y., Dong, R., Yu, L., 2020. Cross-regional oil palm tree counting and detection via a multi-level attention domain adaptation network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167, 154–177.

Zhu, J.-Y., Park, T., Isola, P., Efros, A. A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE International Conference on Computer Vision (ICCV),*.