Multi-Task Learning of Height and Semantics from Aerial Images

Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Frédéric Champagnat, Andrés Almansa

Multi-Task Learning of Height and Semantics from Aerial Images

Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Frédéric Champagnat and Andrés Almansa

***Abstract*—Aerial or satellite imagery is a great source for land surface analysis, which might yield land use maps or elevation models. In this investigation, we present a neural network framework for learning semantics and local height together. We show how this joint multi-task learning benefits to each task on the large dataset of the 2018 Data Fusion Contest. Moreover, our framework also yields an uncertainty map which allows assessing the prediction of the model. Code is available at https://github.com/marcelampc/mtl_aerial_images .**

***Index Terms*—Multi-task learning, Aerial imagery, Semantic Segmentation, Single view depth estimation, Neural networks, Deep learning**

I. Introduction

Aerial imagery has never been so common, even at Very High Resolution (VHR), now that everyone can access images from around the world in any computer. Its automatic analysis is also in progress and has been boosted in the last decade by the tremendous progresses of neural network models. It includes spectral analysis, change detection, and two applications which are of particular interest in this study: semantic mapping of the land surface and local height estimation.

Adding semantics to images by creating high-quality land-cover maps is crucial for environment analysis or urban modelling. A standard way to formulate this problem is classification of each pixel, now reframed as semantic segmentation [1], [2]. Besides, providing the local height in the form of Digital Surface Models (DSMs) is useful for urban planning, telecommunications, aviation, and intelligent transport systems. It has been traditionally done by multi-view stereo [3] until that recently, deep learning approaches also offer competitive performances [4]. Eventually, [5] made one step further by combining both tasks through multi-task learning. In part II, we explore more related work to give insight into these approaches.

We also tackle this problem by using a multi-task deep network that simultaneously estimates both height and semantic maps from a single aerial image *and we show that both tasks benefit from each other*. To reach this goal, our approach builds on powerful models for depth prediction from a single image [6]. Moreover, we investigate the model uncertainty [7] to bring a new regard to aerial imagery processing and better understand success and failure cases. We also perform a study with most recent multi-task techniques for scene-understanding to analyse their contributions when confronted to the DFC2018 dataset. Notably, we obtain state-of-the-art results using Very-High-Resolution (VHR) imagery only on reference datasets: IEEE GRSS DFC2018 [8] and ISPRS Semantic Labeling [9]. On challenging DFC2018 data, comparing true deep learning methods only, our approach achieves 8% more accuracy than the winning solution Fusion-Net without post-processing [10].

After the related work in part II, we will describe our multi-task approach in part III and present results in part IV.

M. Carvalho, B. Le Saux, P. Trouvé-Peloux and F. Champagnat are with DTIS, ONERA, Université Paris-Saclay, FR-91123 Palaiseau, France

A. Almansa is with Université Paris Descartes, FR-75006 Paris, France

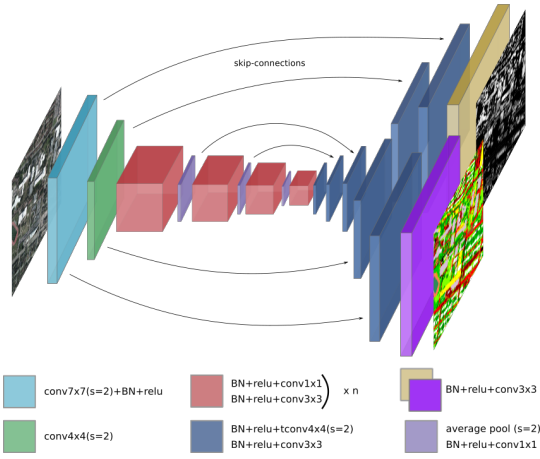

Fig. 1. Architecture of our multi-task model for aerial imagery, based on D3-Net [6]. Left most layers share parameters between all tasks and right most layers are task-specific. Last layer of each decoder differs only on the output number of channels, followed by task evaluation metric (cross-entropy, or $\mathcal{L}_1$, for semantic and height estimation, respectively).

II. Related work

a) Semantic segmentation: this task consists in giving a class label to each pixel in the image [11] and has been commonly carried out in the recent years by Fully-Convolutional Networks (FCNs) since [12]. In remote sensing, it corresponds to the old problem of land-surface classification [13] and has been popularized again by recent benchmarks on urban land-use mapping [9], [8]. Current state-of-the-art approaches based on FCNs include [1], [2] or [14] which combines segmentation with boundary detection. When multi-source data is available, as in the 2018 DFC, dedicated network architectures such as Fusion-CNN [8] can be designed to use this information.

b) Elevation / depth estimation: the problem at hand here is to estimate the distance between the sensor and the observed scene, which means depth in computer vision or elevation (up to an affine transformation) in remote sensing. For nearly 40 years, stereo or multi-view stereo have been the means of

choice, and still give excellent results [3]. However, recently 3D estimation from a single image became popular thanks to the conjunction of availability of image-and-depth datasets with finally powerful-enough neural network models [15], [16]. In remote sensing, several networks for predicting elevation were also proposed, first [5] then [17], [18], [19]. In particular, [19] uses a ResNet-based FCN to produce the DSM while [18] adds an adversarial loss to improve the likelihood of the synthesized DSM. Some works show how the estimated DSM is an useful additional information for building detection [17] or semantic segmentation [18]. It is worth noting that the 2019 Data Fusion Contest [4] comprises one challenge about *Single-view Semantic 3D Challenge* which should yield to new methods to tackle this problem in remote sensing. With respect to previous methods, our approach jointly learns height and semantics, using only VHR imagery.

c) Multi-task Learning: It aims at discovering the latent relatedness among tasks to improve generalization. In practice, it leverages the domain-specific information contained in the training signals of related tasks to build a better model which benefits all tasks [20]. Recent works include the simultaneous prediction of depth, normals and semantic labels [15] or normalized DSM and semantic labels [5]. In the latter, the network consists mostly in shared hidden convolutional layers followed by task-specific heads: one fully-connected layer and the appropriate loss. With respect to theirs, our multi-task architecture favors a middle split for the division in two task-specific branches, a more suitable strategy for tasks as diverse as semantic mapping and DSM regression as this gives the model more specialized layers for each objective. As Caruana pointed out, backpropagation is one of the mechanisms to discover task relatedness [20]. Several recent works have dealt with the balance of the influence of each task during backpropagation. They weight task specific losses according to their intrinsic uncertainty [21], directly the gradient magnitude for GradNorm [22], or to Pareto improvements between the conflicting tasks for MTL-MGDA [23].

III. Method

Network Architecture. We adapted D3-Net, an encoder-decoder deep network originally created for depth estimation, to a multi-task architecture by adding a semantic classification decoder. This architecture favors hard parameter sharing: as illustrated in Fig. 1, the contractive and the early decoder layers are common for both semantics and height estimation. Last layers of the decoders are specific for each objective and generate respectively as many channels as classes for semantics and one channel for height. We also implemented various mechanisms for balancing tasks during optimization following literature in [22], [23]

Multi-Objective Loss. As discussed in section II, learning multiple tasks requires to correctly balance each objective's contribution at every training iteration. Indeed, each output is evaluated with a corresponding loss function: we adopt the absolute error ($\mathcal{L}_1$) for height regression and the cross entropy loss ($\mathcal{L}_{\text{CE}}$) for semantics evaluation. Thus, when errors are backpropagated, the resulting gradient in the common layers will correspond to the sum of all task gradients. A simple way to control the contribution of each task consists in multiplying each loss term in the final loss 1 by a scalar k_t. However, finding the optimal values for each k_t is still challenging.

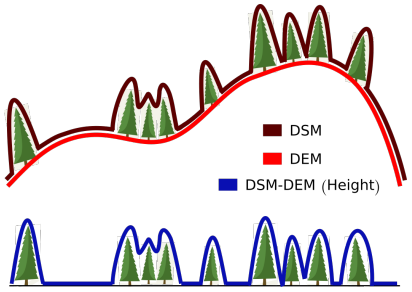

Fig. 2. Height map generation using the Digital Surface Model (DSM) and the Digital Elevation Model (DEM): height as DSM-DEM.

$$L_{final} = \sum_{t=1}^{T} k_t \mathcal{L}_t \tag{1}$$

Consequently, many methods have been proposed to effectively estimate k_t in order to converge common parameters values to the best model for all tasks. Here, we propose to evaluate these methods in the context of aerial imagery. In the following, we explain the main idea of each approach.

a) Equal weights: is the most common approach and consists on weighting all losses uniformly. This approach does not handle cases when the training errors have different scales. In consequence some tasks can be dominant and predictions may be degraded for the other ones. However, this technique can still be effective in the case we can not appropriately measure the best contribution of each task to the global model.

b) GradNorm [22]: dynamically learns the scaling factors with respect to the gradients of the last common layer and the rate balance, defined as the relative inverse training rate for each task. By directly modifying gradient magnitudes with learnable parameters, this method does not rely on empirical values for k_t.

c) MTL-MGDA [23]: Sener et al. proposed to adapt the Multiple Gradient Descent Algorithm (MGDA) to the multi-objective optimization. This approach uses the Frank-Wolfe algorithm [24] to find a common descent direction to the gradients of the shared layers at each iteration. Besides that, the paper also proposes to reduce memory use by applying the MGDA to the upper-bound *MTL-MGDA-UB* of the objective by means of the gradient of task losses with respect to the intermediate representation on the last common layer.

IV. Experiments

In this section, we first describe the two datasets and the metrics used in this study. Experiments are as follows. First, we compare the general approach of multi-task using equal weights to single-task objectives and also to current state-of-art results on the proposed datasets to validate our approach and architecture. Then, we extent our study by analyzing the uncertainty map of the proposed model, according to [25]. Finally, we compare various flavors of recent state-of-art multi-task in this context.

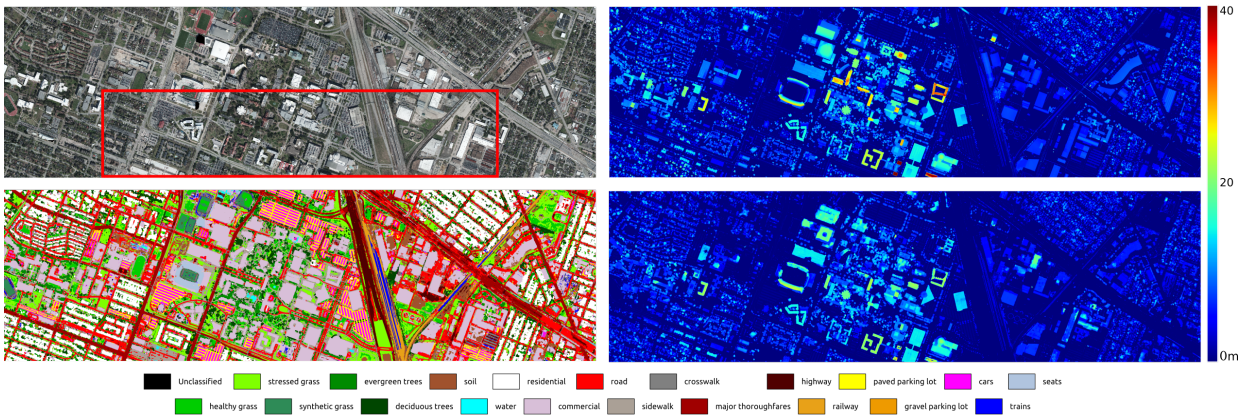

Fig. 3. Results on the DFC2018 dataset trained with equal weights (best results). Top row shows RGB image and height ground-truth, bottom row semantic prediction and height estimate. The training area is delimited by a red rectangle on the RGB image.

A. Datasets and Metrics

The **ISPRS Vaihingen** [9] dataset comprises IRRG (Infra-Red, Red and Green) images at 9cm / pixel, DSM, and 2D and 3D semantic labeled maps for urban classification and 3D reconstruction. It contains 33 patches of different sizes, of which 16 images are used for training and the remaining 17 are used for testing. Semantic maps were annotated with 6 classes including impervious surfaces, building, low vegetation, tree, car and clutter/background. We ignore this last class during training and testing. As in [5], for height estimation we adopt the normalized DSMs (nDSM) from [26].

The **2018 Data Fusion Contest** (DFC2018 [8]) dataset is a collection of multi-source optical imagery over Houston, Texas. In particular, it contains Very High Resolution (VHR) color images resampled at 5cm / pixel, hyperspectral images, and LiDAR-derived products such as DSMs, and Digital Elevation Models (DEMs) at a resolution of 50cm / pixel. A 20-class, handmade ground-truth exists: 4 tiles (corresponding to the VHR images in the red frame in Fig. 3) are available for training, while 10 tiles remain undisclosed for evaluation on the the DASE website [1]. The original DFC2018 dataset does not include height maps originally, thus we generate them by substracting the DEM to the DSM, as illustrated in Fig. 2.

Pre-processing. For both datasets, we perform training using 320x320 crops from the original images. As mentioned, RGB images from DFC2018 are 10 times bigger than the height and semantic models. So, we perform training with two different strategies: first, to deal with VHR images, we upsample the height and semantic maps to the same resolution of the input image before performing crops; second, to speed up training and testing, we downsample the RGB images by a factor of 10. We refer to these pre-processing strategies as VHR Multi-task and LR (low-resolution) Multi-task.

Data augmentation. To improve generalization, we perform the following online data augmentation: random crops from original tiles, rotation, horizontal and vertical flips.

From crops to tiles. Inference is implemented using a Gaussian prior over patches to avoid a checkerboard effect on the output. We predict patches sequentially with a stride smaller than the window size and weight overlapping areas with a 2D Gaussian map. Results are improved when using bigger windows and small strides as we can leverage more information from neighbor patches. For our experiments, we use a test window of 1024 and a stride of 256. When generating VHR outputs, these are posteriorly downsampled to compare to ground truth maps.

Training is performed with PyTorch [27] framework, we used Adam [28] as our optimizer with learning rate of 2e-4 and we train our model with a Nvidia GTX 1080 GPU.

Metrics. To evaluate our models, we use common metrics from [15] and [5]. For height estimation, we use the mean absolute error (mae): $\frac{1}{N}\sum_{i=0}^{N}|d_i - \hat{d}_i|$; the mean squared error (mse), $\frac{1}{N}\sum_{i=0}^{N}\left\|d_i - \hat{d}_i\right\|^2$; and the root mean squared error (rmse), $\sqrt{\frac{1}{N}\sum_{i=0}^{N}(d_i - \hat{d}_i)^2}$. For classification, we use overall accuracy (OA), average accuracy (AA) and Kappa, as in [8].

TABLE I
COMPARISON OF DFC2018 HEIGHT AND SEMANTICS PREDICTIONS WITH STATE-OF-ART APPROACHES, SINGLE AND MULTI-TASK MODELS.

	Height Errors			Semantic Errors			Time[1](s)
	mae↓	mse↓	rms↓	OA(%)↑	AA↑	Kappa↑	
Cerra* [29]	-	-	-	58.60	55.60	0.56	-
Fusion-FCN* [10]	-	-	-	63.28	-	0.61	-
Fusion-FCN [10]	-	-	-	80.78	-	0.80	-
*learning only, without post-processing							
VHR Single task	1.480	9.544	3.000	73.40	67.82	0.72	-
VHR Multi-task	1.263	7.279	2.599	74.44	68.30	0.73	7.74.10[2]
LR Multi-task	1.513	9.341	2.970	64.70	58.85	0.63	7.82

[1]Mean test time per image (std. deviation)

B. Improving Height and Semantic Estimation with Multi-task

In this experiment, we use the original D3-Net with the corresponding task decoder for single task training, and the proposed model with *equal weights* for multi-task.

If we focus on the bottom lines of Table I, we observe that performances are improved for both objectives by the multi-task model if compared to the single-task. It has the advantage of learning complementary features, using less parameters compared to single models for each task.

[1]GRSS Data and Algorithm Standard Evaluation website: http://dase.grss-ieee.org/

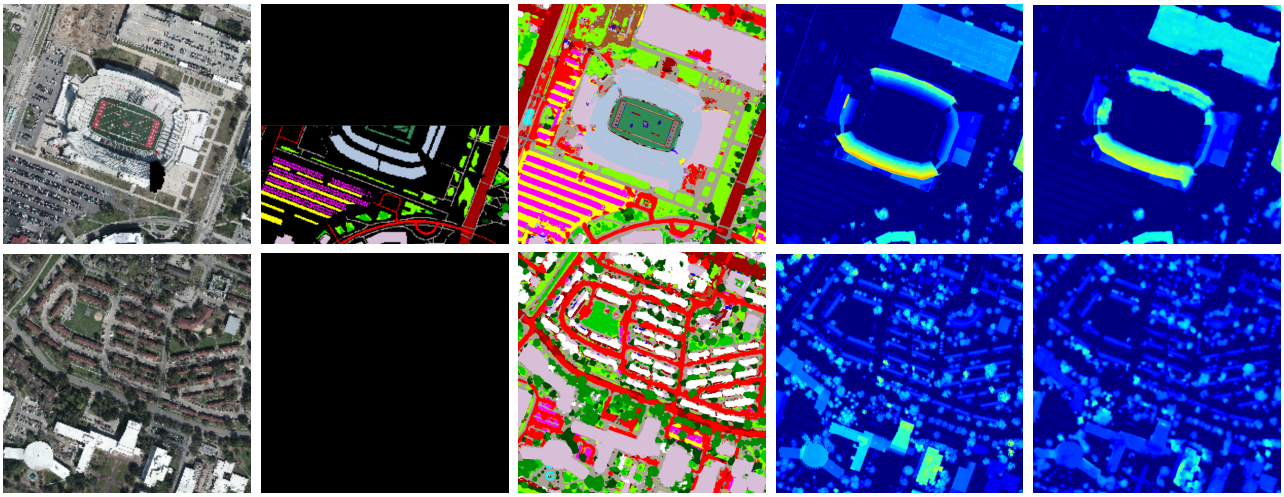

Fig. 4. Crop areas over the DFC2018 dataset. From left to right, input RGB image, semantic ground-truth and prediction (black is no information), height ground-truth and prediction. Top row show the Houston University stadium and bottom row shows a residential area.

TABLE II
COMPARISON OF MULTI-TASK STATE-OF-ART AND OUR ARCHITECTURE ADOPTING AN EQUAL WEIGHT APPROACH.

	Height Errors			Semantic Errors		
	mae↓	mse↓	rms↓	OA*(%)↑	AA**↑	Kappa↑
Srivastava [5]	0.063	-	0.098	78.8%	73.4%	71.9%
Single task	0.039	0.005	0.067	87.4%	84.4%	75.0%
Multi-task	0.045	0.006	0.074	87.7%	85.4%	75.9%

In the upper lines, we observe state-of-art results for DFC2018. Please note these methods use multi-source data from DSM, DEM, Hyperspectral image and VHR RGB as input to estimate semantic maps only. For fairness of comparison, results with a $*$ refer to methods without ad-hoc detector nor post-processing.It appears our model overcomes past learning-based approaches by 10 percentual points on OA.

The above results are inferred in nearly 13 minutes for each $10^4 \times 10^4$pixels tile, when using the inference proposed in section III. For large batch processing, this time can even be reduced by using the LR model, which reduces time to nearly 10 seconds per tile at the cost of losing efficiency.

We can also observe results in Fig. 3 and crops for specific regions in Fig. 4. In general, the network produces nearly accurate results for ground, residential buildings and vegetation, while some structures are more challenging, like high buildings or stadiums. These classes have various shapes, colors and heights. Thus, it is hard to estimate precise height values from bird-view. Semantics are detailed, with even plastic seats, playground or concrete elements in the stadium.

Results with Vaihingen dataset are in Table II. We observe that our performances overcome Srivastava et al. [5]. This is likely due to a better network with skip-connections and an earlier split between task-specific decoders. It is worth noting that our multi-task approach only improves semantic classification if compared to single-task models. In [5], none of the tasks was improved by multi-task. Possible reason is that Vaihingen does not have much variance between train and test sets, so even single task models can overfit and also be performing during inference. Actually, multi-task improves generalization by leveraging complementary information between tasks.

C. Uncertainty

In addition to error measures, [7] proposed to evaluate the uncertainty of the network, which accounts for the ignorance of the model parameters with respect to the input images. To perform this analysis, we follow the original paper and keep dropout layers active during inference. For each tile, we generate 30 samples from which we calculate the standard deviation of the predictions. We perform this test for height estimation only. The results in Fig. 5 allows us to understand which zones of the input are the most challenging to the network. Contours present high variance and are challenging. So are high buildings in general: indeed a plane rooftop appears the same whatever its altitude. We also note that trees are quite uncertain even if predictions were good: this is a difficult class due to texture variance or deciduousness.

D. Comparison between Multi-task Methods

In this section, we compare the classic approach with equal weights to state-of-art methods for multi-task learning. These techniques were previously tested on datasets for digit classification, multi-label classification, urban outdoor and indoor scene understanding (Cityscape [30] and NYUv2 [31]). We now test them for the first time on VHR aerial images.

As the chosen methods originally rely on architectures without skip connections, we perform experiments with and without these features for best comparison. We observe results for the mentioned methods in Tables III and IV.

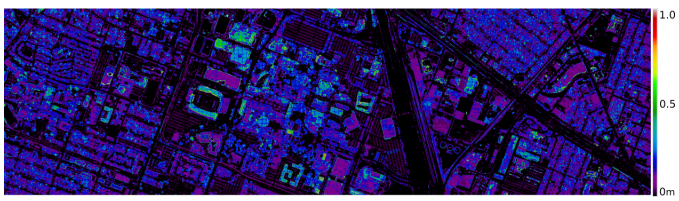

Fig. 5. Uncertainty map of height (standard deviation of model predictions).

From our experiments, we observed that aerial imagery requires less context for the objective tasks than scene-parsing datasets in [22], [23]. We believe that subtle multi-task techniques are more prone to better results on these kind of datasets,. Also, in the case of DFC2018, semantic annotations are very sparse and gradient values are impacted, which compromising other multi-task methods.

TABLE III
COMPARISON OF DIFFERENT MULTI-TASK APPROACHES FROM THE STATE-OF-ART WITH VAIHINGEN DATASET

		Height Errors↓			Semantic Errors↑		
	MTL Method	mae	mse	rms	OA(%)	AA(%)	Kappa
no skip	MTL-MGDA-UB [23]	0.042	0.006	0.075	81.9%	66.0%	55.8%
	GradNorm [22]	0.044	0.006	0.074	87.3%	84.2%	74.3%
	Equal Weights	0.047	0.006	0.076	87.3%	84.2%	74.6%
skip	MTL-MGDA [23]	0.042	0.007	0.079	85.8%	81.4%	71.2%
	GradNorm [22]	0.040	0.005	0.068	87.4%	85.1%	75.4%
	Equal Weights	0.043	0.006	0.073	87.5%	84.9%	75.5%

TABLE IV
COMPARISON OF DIFFERENT MULTI-TASK STATE-OF-ART APPROACHES WITH VHR INPUT IMAGES FROM DFC2018 DATASET.

		Height Errors↓			Semantic Errors↑		
	MTL Method	mae	mse	rms	OA(%)	AA(%)	Kappa
no skip	MTL-MGDA-UB [23]	1.475	9.911	3.047	52.98	47.59	0.50
	GradNorm [22]	1.394	8.886	2.857	58.00	54.23	0.56
	Equal Weights	1.520	8.589	2.826	58.26	54.74	0.56
skip	MTL-MGDA [23]	1.303	7.415	2.627	59.13	55.53	0.57
	GradNorm [22]	1.340	7.898	2.743	63.07	58.92	0.61
	Equal Weights	1.263	7.279	2.599	74.44	68.30	0.73

V. CONCLUSIONS

In this work, we have shown that multi-task learning methods work really well on aerial imagery and may lead to better results when compared to single-task techniques. We proved that complementary features from each objective can be learned by a deep model to improve performance independently. Our experiments on DFC2018 show that a model with less input data and no special post-processing can lead to results comparable to the much complex state-of-art results. Thus, this framework can be easily adopted for urban modelling without the any complementary information. However, experiments with very recent multi-task variants showed that surprisingly the simple equal-weight approach leads to best performances. Maybe subtle multi-task methods require larger and densely labeled datasets. Hence, we will further our work to understand the mechanisms of multi-task.

REFERENCES

[1] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van Den Hengel, "Effective semantic pixel labelling with convolutional networks and Conditional Random Fields," in *Proc. CVPR/W*, Hawaii, USA, 2015.

[2] N. Audebert, B. Le Saux, and S. Lefevre, "Beyond RGB: Very High Resolution Urban Remote Sensing With Multimodal Deep Networks," *ISPRS J. Photogram. and Remote Sensing*, 2018.

[3] G. Facciolo, C. de Franchis, and E. Meinhardt-Llopis, "Automatic 3D reconstruction from multi-date satellite images," in *Proc. ICCV Workshops*, Santiago, Chile, 2015.

[4] B. Le Saux, N. Yokoya, R. Hansch, M. Brown, and G. Hager, "2019 IEEE GRSS Data Fusion Contest: Large-Scale Semantic 3D Reconstruction," *IEEE Geosci. and Remote Sensing Mag.*, 2019.

[5] S. Srivastava, M. Volpi, and D. Tuia, "Joint height estimation and semantic labeling of monocular aerial images with CNNs," in *Proc. IGARSS*, Fort Worth, USA, 2017.

[6] M. Carvalho, B. Le Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat, "On regression losses for deep depth estimation," in *Proc. ICIP*, Athens, Greece, 2018.

[7] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[8] Y. Xu, B. Du, L. Zhang, D. Cerra, M. Pato, E. Carmona, S. Prasad, N. Yokoya, R. Hänsch, and B. Le Saux, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS Data Fusion Contest," *IEEE J. Sel. Topics in Applied Earth Obs. and Remote Sensing*, vol. 12, no. 6, 2019.

[9] M. Cramer, "The DGPF-test on digital airborne camera evaluation–overview and test design," *Photogrammetrie-Fernerkundung-Geoinformation*, 2010.

[10] Y. Xu, B. Du, and L. Zhang, "Multi-source remote sensing data classification via fully convolutional networks and post-classification processing," in *Proc. IGARSS*, Valencia, Spain, 2018.

[11] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Patt. Rec. Lett.*, 2009.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proc. CVPR*, 2015.

[13] J. Benediktsson, P. H. Swain, and O. Ersoy, "Neural network approaches versus statistical methods in classification of multisource remote sensing data," in *Proc. IGARSS*, Vancouver, Canada, 1989.

[14] D. Marmanis, K. Schindler, J. Dirk Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogram. and Remote Sensing*, 2016.

[15] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. ICCV*, Santiago, Chile, 2015.

[16] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 3DV*, Stanford, USA, 2016.

[17] L. Mou and X. X. Zhu, "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," *arXiv preprint arXiv:1802.10249*, 2018.

[18] P. Ghamisi and N. Yokoya, "IMG2DSM: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geosci. and Remote Sensing Lett.*, 2018.

[19] H. A. Amirkolaee and H. Arefi, "Height estimation from single aerial images using a deep convolutional encoder-decoder network," *ISPRS J. Photogram. and Remote Sensing*, 2019.

[20] R. Caruana, "Multitask learning," *Machine Learning*, 1997.

[21] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. CVPR*, Salt Lake City, USA, 2018.

[22] Z. Chen, V. Badrinarayanan, C. Lee, and A. Rabinovich, "GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. ICML*, Stockholm, Sweden, 2018.

[23] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Proc. NeurIPS*, Montréal, Canada, 2018.

[24] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *Proc. ICML*, Atlanta, USA, 2013.

[25] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *arXiv preprint arXiv:1703.04977*, 2017.

[26] M. Gerke, "Use of the Stair Vision Library within the ISPRS 2D semantic labeling benchmark (Vaihingen)," *Techical Report*, 2015.

[27] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. NeurIPS Workshops*, Long Beach, USA, 2017.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, Banff, Canada, 2014.

[29] D. Cerra *et al.*, "Combining deep and shallow neural networks with ad hoc detectors for the classification of complex multi-modal urban scenes," in *Proc. IGARSS*, Valencia, Spain, 2018.

[30] M. Cordts *et al.*, "The Cityscapes dataset for semantic urban scene understanding," in *Proc. CVPR*, Las Vegas, USA, 2016.

[31] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. ECCV*, Firenze, Italy, 2012.