Multitask Learning for Human Settlement Extent Regression and Local Climate Zone Classification

Chunping Qiu, Lukas Liebel, *Graduate Student Member, IEEE*, Lloyd Haydn Hughes, *Member, IEEE*,
Michael Schmitt, *Senior Member, IEEE*, Marco Körner, *Member, IEEE*,
and Xiao Xiang Zhu, *Senior Member, IEEE*

Abstract—Human settlement extent (HSE) and local climate zone (LCZ) maps are both essential sources, e.g., for sustainable urban development and Urban Heat Island (UHI) studies. Remote sensing (RS)- and deep learning (DL)-based classification approaches play a significant role by providing the potential for global mapping. However, most of the efforts only focus on one of the two schemes, usually on a specific scale. This leads to unnecessary redundancies since the learned features could be leveraged for both of these related tasks. In this letter, the concept of multitask learning (MTL) is introduced to HSE regression and LCZ classification for the first time. We propose an MTL framework and develop an end-to-end convolutional neural network (CNN), which consists of a backbone network for shared feature learning, attention modules for task-specific feature learning, and a weighting strategy for balancing the two tasks. We additionally propose to exploit HSE predictions as a prior for LCZ classification to enhance the accuracy. The MTL approach was extensively tested with Sentinel-2 data of 13 cities across the world. The results demonstrate that the framework is able to provide a competitive solution for both tasks.

Index Terms—Convolutional neural networks (CNNs), human settlement extent (HSE), land cover (LC), local climate zones (LCZs), multitask learning (MTL), sentinel-2.

I. INTRODUCTION

TWO important tasks in urban mapping are distinguishing urban areas from nonurban background and characterizing intraurban heterogeneity. Human settlement extent (HSE) and local climate zone (LCZ) are two schemes for the respective representations. HSE density, indicating the portion of buildings, roads, and other man-made structures in an area (e.g., a pixel), depicts the human footprint on Earth. By contrast, the LCZ scheme, originally proposed for Urban Heat Island (UHI) studies, includes 17 classes for detailed land cover (LC) classification [1] and is universally applicable. Up until date, detailed and accurate worldwide information on HSE and LCZ can provide support for evidence-based decision making for various applications, e.g., global climate science [2], [3].

The deep learning (DL)-based approaches for LC classification have been attracting much attention due to their proven predictive power and end-to-end learning abilities of complex feature representations. Convolutional neural network (CNNs), in particular, fully convolutional network (FCN), have been successfully applied to many remote sensing (RS)-based image classification and segmentation tasks [4]. In this study, we draw upon these successes by applying DL-based semantic segmentation methods to LC mapping. However, LC mapping, particularly LCZ classification over a large scale, remains challenging due to the large intraclass variability of spectral signatures, which stems from variations in physical and cultural environmental characteristics across the world and a limited amount of reference data [5].

This challenge and the lack of fully automated approaches to it motivate our investigations into DL models with a higher generalization ability for LC mapping. As already suggested by the definition of the two schemes, prior work also indicated that HSE and LCZ have close correspondence for different study areas [1], [6]. To exploit the complementary nature of the HSE regression and LCZ classification tasks, we propose a multitask learning (MTL) framework to jointly predict HSE and LCZ, considering that MTL has been shown to be a powerful technique for improving model generalization by leveraging domain knowledge of related complementary tasks [7], [8]. In this work, we present a feature-based MTL system that mainly consists of a shared backbone network to capture a common representation for both HSE regression and LCZ classification, and soft-attention modules to adaptively select task-specific features. By jointly learning both tasks,
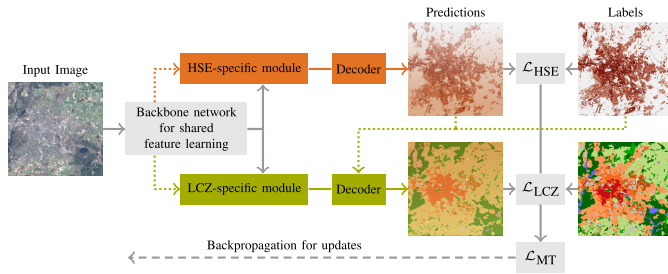
Fig. 1.   General MTL framework for HSE density regression and LCZ classification, consisting of a backbone network, task-specific network branches, and decoder modules. The inputs for network training are images and corresponding reference labels for each task.
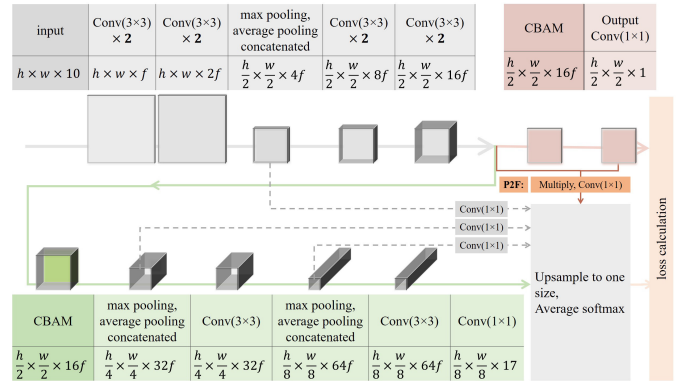

Fig. 2.   Illustration of the implemented MTL CNN architecture for HSE regression and LCZ classification. The backbone network consists of two convolutional blocks, one pooling block, and two more convolutional blocks. The two task-specific network branches are indicated by two different colors. The description of each layer and the size of inputs, feature maps, and outputs are listed along with the operations. h and w are height and width of the input patch, and f is the number of feature maps from the first layer.

we aim to boost the prediction performance over that achieved by a single, task-specific network.

II. METHODOLOGY

A. General MTL Framework for LC Mapping on Different Scales

In the definition of LCZ, one important factor to consider is the ratio of impervious surface and buildings, which indicates that the HSE density affects the categorization of an area in the LCZ scheme. In addition, HSE regression and LCZ classification prioritize spatial and semantic resolution in a complementary manner. The LCZ scheme places emphasis on the characterization of urban morphology with 17 classes (for a relatively large neighboring area, e.g., 100 m), while the HSE density contains fewer semantic details but features a higher spatial resolution (e.g., 20 m). The complementary relationship, when appropriately exploited, is expected to support the learning process of each task and improves the efficient usage of available training data. This will potentially lead to faster convergence during training, higher accuracy for both tasks compared with single-task models, and reduced production time as both tasks can occur in parallel. To this end, a generalized MTL framework, as illustrated in Fig. 1, is proposed, which consists of the following four primary components.

1) A backbone network for shared, general feature learning at the early stage. It can be implemented as state-of-the-art CNN architectures, such as residual neural networks. Compared with natural images, e.g., photographs typically used in computer vision experiments, the resolution of Sentinel-2 imagery is low. Hence, pooling should be less used within the backbone network to retain the spatial resolution and the information of the pixel values.

2) Task-specific network branches for dedicated feature learning to adaptively select features from the common representations and further learn specific and high-level features for the individual tasks. They can be implemented as convolutional operations, attention modules, or "Squeeze-and-Excitation" blocks [9].

3) Decoder modules to recover the required resolution of predictions. They can be implemented as transposed convolutions or upsampling layers. To retain clear boundaries, low-level features might need to be considered, e.g., via skip connections.

4) Modules to exploit task relation. The reference of one task can be employed as prior information to guide the prediction of the other task. In this way, consistent

and accurate predictions are encouraged in the MTL framework.

B. Implementation of the MTL Framework

One implementation of the proposed MTL framework is illustrated in Fig. 2, which consists of a backbone network (including convolutional layers and pooling layers) followed by two branches for HSE regression and LCZ classification. Each of the two branches begins with an attention module, implemented as a convolutional block attention module (CBAM), for an adaptive selection and learning of task-specific representations [10]. All 3×3 convolutional layers make use of separable convolution operations for the sake of efficiency. A kernel size of 2×2 with a stride of 2 was used for the pooling layers, decreasing the size of feature maps by half. Maximum and average pooling layers were used together for abstracting learned features within the architecture to preserve sufficient features. The final output of both tasks was decided based on the desired ground sampling distance (GSD) of the respective products. To avoid overfitting during training, one and two dropout layers with a dropout rate of 0.1 were utilized in the HSE and LCZ branches (omitted in the illustration).

For the HSE regression task, the last layer was activated with a sigmoid function, and the mean absolute error (MAE) was used as a loss function to consider potential noise in the reference data. Considering that there are more samples with no or few human settlements in our data set, a relatively high weight was assigned to samples with dense human settlements to deal with the imbalance problem. The sample weight was decided based on: $e^{y_{HSE}}$, where $y_{HSE} \in [0, 1]$ is the reference label for HSE density.

For the LCZ classification task, predicted softmax probabilities from intermediate lower level features were also used for the loss calculation, in addition to the final prediction. Especially, for each input patch, three 1×1 convolutional and softmax layers were used to independently predict three results using intermediate features (as indicated by dashed gray lines in Fig. 2). This is to fully exploit the features at different levels for the elaborate LCZ scheme, which requires a diversity of representations to distinguish the 17 distinct LCZ classes.

Together with the result produced by the final layers (represented by a solid green line in Fig. 2), these four results were upsampled to the same size as the prepared reference label patches. All four patches of predicted softmax probabilities were averaged into one final patch for the loss calculation of the LCZ branch using a softmax cross-entropy loss.

The final MTL loss, which was used to train the MTL network, was a combination of the two single-task losses from the HSE regression task and the LCZ classification tasks.

C. Predictions as Features (P2F)

The HSE density and LCZ do not merely share similar properties, which motivated the implicit exploitation via an MTL framework. The HSE density is, furthermore, explicitly affecting the categorization of a subset of LCZ. For instance, an area with a high HSE density cannot belong to an LCZ class that corresponds to natural areas, e.g., dense trees. Following this principle, a module called predictions as features (P2F) was designed to exploit this prior knowledge within our framework, the implementation of which was inspired by Kohl *et al.* [11]. The main idea is to use the HSE predictions as a prior for the classification of LCZ, as illustrated in Figs. 1 and 2. Especially, the HSE reference was multiplied with the intermediate features, resulting in processed feature maps. These feature maps were used to get an additional prediction of LCZ classes, providing an additional output of the LCZ branch.

The P2F module was used in a different manner during training and test time. Since reference labels for HSE were available at the training time, they can be utilized as the prior for the prediction of LCZ. At the test time, the prior was predictions of the HSE regression branch in the multitask network, as indicated in Fig. 1. Thus, the system still solely relies on a single image as its input at the test time while making use of the available ground-truth data during training. The quality of HSE predictions is expected to be useful for the P2F concept as satisfying HSE mapping results have been achieved using deep CNN [12].

D. Dynamically Balancing Task Weights

One of the challenges in MTL is to balance the involved tasks, which is most commonly implemented by weighting each individual loss in the multitask loss, subject to optimization. Simply summing the contributing losses, assuming equal task weights of 1, is not sufficient in most cases, as this not only implies equal importance of the tasks but also that single-task losses produce values in the same order of magnitude. Manually tuning such weights as hyperparameters, however, is tedious. Hence, approaches to automatically and dynamically determine them are desired and have been investigated in the literature [13], [14].

We implemented an approach to weighing the HSE regression and LCZ classification tasks based on homoscedastic uncertainty, as introduced by Kendall *et al.* [13]. Especially, we optimize a multitask loss function

$$\mathcal{L}_{MT}(\mathbf{W}_{HSE/LCZ}, \sigma_{HSE/LCZ}) = \frac{1}{2\sigma_{HSE}^2}\mathcal{L}_{HSE}(\mathbf{W}_{HSE})$$
$$+ \frac{1}{\sigma_{LCZ}^2}\mathcal{L}_{LCZ}(\mathbf{W}_{LCZ})$$
$$+ \log\sigma_{HSE} + \log\sigma_{LCZ} \quad (1)$$

where $\mathcal{L}_{HSE/LCZ}$ is the MAE and cross-entropy loss for HSE regression and LCZ classification task, respectively, $\mathbf{W}_{HSE/LCZ}$ are the trainable parameters from the network layers of respective branch, and $\sigma_{HSE/LCZ}$ are weighting parameters that control the contribution of the individual tasks. The variables for the input and reference data, $\mathbf{X}$ and $\mathbf{Y}$, are omitted in the above equation. The regularization terms $\log\sigma_{HSE/LCZ}$ prevent trivial solutions for $\sigma_{HSE/LCZ} \to \pm\infty$. We optimized the weighting terms along with the network parameters as $s_{HSE/LCZ} := \log\sigma_{HSE/LCZ}^2$ due to numerical stability [13].

III. EXPERIMENTAL EVALUATION

A. Data Set for Training

A data set was prepared to assess the potential of the proposed MTL framework, including Sentinel-2 image patches from three seasons (spring, summer, and autumn 2017) and annotations for both tasks, i.e., HSE density percentage and LCZ labels.

The imagery was processed in accordance with previous work [12] using the five European cities, namely, Berlin, Lisbon, Madrid, Milan, and Paris, as study areas. Annotations for the HSE density regression are from "High Resolution Layer Imperviousness 2015," using continuous values indicating the percentage of each pixel covered by HSE, with a GSD of 20 m [15], [16]. In addition, pixelwise LCZ labels were prepared for each sample, using the reference from the World Urban Database and Access Portal Tools (WUDPAT) project [17], with an original GSD of 100 m. It was upsampled to 10 m to match the coregistered image patches during the training period. At test time, LCZ predictions with a GSD of 100 m were produced to be consistent with state-of-the-art LCZ-related studies. The number of spatially disjoined training and validation patches was 75 116 and 24 706, respectively, with a size of 128×128 px.

B. Test Data and Accuracy Assessment

For assessment of the HSE regression task, two test data sets were used. The first one covers three test scenes in Europe, namely, Amsterdam, London, and Munich, and was prepared in the same way as the training data. Hence, the HSE density data are available for this test data set. The MAE is used as a metric when tested on this data set. To test the HSE regression performance on a large scale, a qualitative evaluation in comparison to high resolution (HR) images was carried out. Furthermore, the predicted results were aggregated into binary labels and tested against manually annotated ground checking points, as introduced by Qiu *et al.* [12], which are uniformly distributed over ten scenes across the world. Metrics for evaluation include Kappa, average accuracy (AA), recall, and F-Score of HSE. This evaluation procedure, with two test data sets, can provide an estimation of the generalization ability of the proposed MTL approach, which was trained on data from Europe exclusively.

To assess the LCZ classification task, we utilized the same ten worldwide scenes as the HSE regression, with reference labels from the So2Sat LCZ42 data set, which was processed by polygon shrinking and class balancing after labeling [18]. The number of test samples for each scene is shown in Fig. 3. In addition to overall accuracy (OA), Kappa,

TABLE I

COMPARISON OF RESULTS FROM DIFFERENT APPROACHES TESTED WITHIN AND OUTSIDE EUROPE. THE VALUES WERE AVERAGED OVER THREE AND TEN TEST AREAS FOR THE EUROPEAN AND THE GLOBAL TEST SET. FOR EACH METRIC, THE TOP TWO BEST RESULTS WERE MARKED IN BOLD

| Configuration	European Test Set					Global Test Set							
	HSE Density	HSE Aggregated				HSE Aggregated				LCZ			
	MAE	Kappa	AA	Recall	F	Kappa	AA	Recall	F	OA	Kappa	AA	WA
Single-Task HSE	4.24	0.76	**0.92**	0.91	0.94	0.78	0.89	**0.93**	0.89	—	—	—	—
Single-Task LCZ	—	—	—	—	—	—	—	—	—	0.48	0.41	0.34	0.85
Multi-Task (1:1, CBAM)	4.55	0.69	0.90	0.86	0.91	0.77	0.89	**0.94**	0.89	**0.50**	0.42	0.34	0.88
Multi-Task (1:1, CBAM, P2F)	4.39	0.71	0.90	0.87	0.92	0.78	0.89	0.92	**0.90**	0.44	0.37	0.31	0.82
Multi-Task (learned weights, CBAM)	4.49	0.76	0.91	0.91	0.94	**0.80**	0.90	0.92	**0.90**	0.48	**0.42**	**0.36**	**0.88**
Multi-Task (learned weights, P2F)	**4.14**	**0.77**	**0.92**	**0.92**	**0.95**	**0.81**	0.90	0.92	**0.91**	0.45	0.38	0.32	0.83
Multi-Task (learned weights, CBAM, P2F)	**4.12**	**0.79**	0.91	**0.93**	**0.95**	**0.80**	0.90	0.90	**0.90**	**0.50**	**0.44**	**0.36**	**0.88**
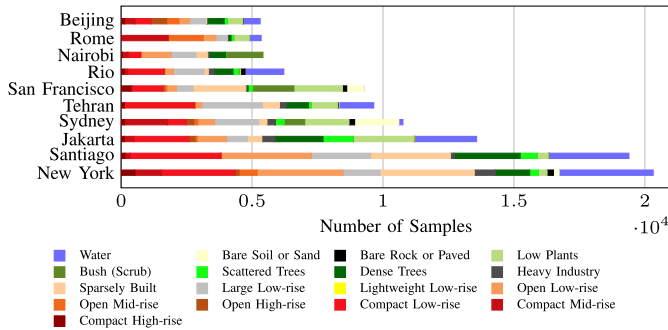
Fig. 3. Sample number of each test scene for LCZ classification assessment.

and AA, weighted accuracy (WA) was used by assigning different penalties to different mistakes.

C. Hyperparameter Settings

All models in this study were implemented in Keras for TensorFlow and trained from scratch. Basic hyperparameters include a batch size of 8 and an initial learning rate of 0.002 for the Nesterov Adam optimizer. The learning rate was decreased by 0.25 after every two epochs. To control the training time and avoid overfitting, early stopping was implemented with the validation loss as the monitored metric with the patience of ten epochs.

D. Results for Different Settings

Table I presents evaluation results for HSE regression and LCZ classification over both test sets. In addition to single-task and multitask trainings, different strategies of feature exploitation and task weighting are also compared.

An illustration of the MTL predictions is presented in Fig. 4 for a test scene in New York City (NYC). Both HSE regression and LCZ classification results reflect the pattern of urban structures in the imagery, indicating a reasonable mapping result.

IV. DISCUSSION

A. Superiority of the Proposed MTL Approach

The joint prediction was able to provide benefits for both tasks, as analyzed in the following. Simply weighing the two contributing single-task losses equally (1 : 1) only resulted in a slight improvement for OA and Kappa in the LCZ classification task, as shown in Table I. When using the dynamically learned weights, the benefit became apparent, with AA being improved from 0.89 to 0.90 and from 0.34 to 0.36 for HSE and LCZ results. This shows that the dynamical weight learning strategy plays a positive role in MTL performance.

The proposed strategy of using the HSE density predictions as features for LCZ classification, P2F, was able to additionally improve the effectiveness of the MTL configuration, providing the best results among all investigated approaches for the LCZ classification task. This is an improvement over the basic approaches, such as single-task learning, with Kappa being improved from 0.41 to 0.44 and OA being improved from 0.478 to 0.50. This effect can be attributed to guidance from the HSE density mapping task. Intuitively, the candidate LCZ type of an area can be narrowed down when built-up areas were already detected in the HSE regression task. This piece of useful information from HSE regression can be leveraged for LCZ classification, leading to a further improvement over baseline MTL approaches. It can also be noted from Table I that P2F was not able to provide improvement when the task weights were 1 : 1, which indicates the importance of an appropriate weighting strategy.

Task-specific feature learning modules, e.g., CBAM utilized in the experiments, are important for gaining benefits within the MTL framework. This can be proved by comparing the results with and without CBAM in Table I. All metrics for LCZ classification were worse when removing CBAM from the MTL architecture. A possible reason is that HSE mapping and LCZ classification each require some different lower level features, which cannot be satisfied in the absence of attention modules as a feature selection process. Another possible reason for the improved results from MTL is that multisource reference data is used. This helps to learn more generalized features, and thus, the model is less prone to overfitting compared with single-task learning.

B. Quality of Current Mapping Results and Future Work

The achieved HSE regression results, representing continuous HSE density, provide richer information beyond a binary delineation of urban areas, such as the global urban footprint (GUF), the global human settlement (GHS) built-up grid, and our previous work [12]. Quantitatively, the accuracy is higher using the same checking points as a test set. Especially, the AA, recall, and F-score of our previous binary HSE
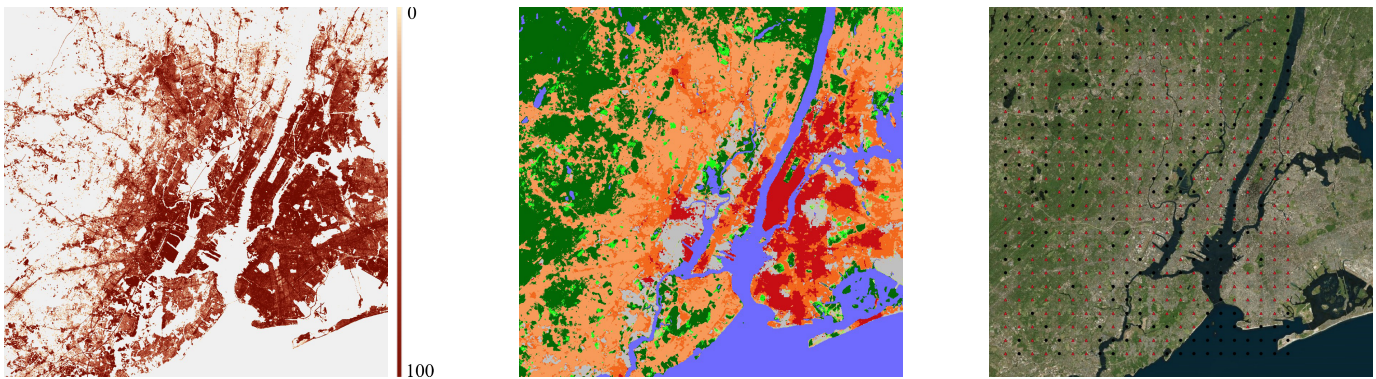
Fig. 4. Illustration of joint prediction results in NYC, USA. From left to right are results of HSE regression and LCZ classification (with the same legend as Fig. 3) and HR image overlaid with reference points (red for HSE and black for non-HSE), respectively.

mapping results are 0.90, 0.91, and 0.91, respectively. All these metrics are beyond the state-of-the-art products when tested on the same ten distinct scenes. For instance, AA of the GUF and the GHS built-up grid are 0.86 and 0.83. It should be noted, however, that the temporal gaps between the checking points and these products might play a role in this comparison.

As a proof of concept, reference data from only five European scenes were collected and used to train the MTL network. Still, the achieved LCZ classification results are promising over ten distinct test scenes across the world, demonstrating the capability for generalization and potential for further exploring this challenging task. The accuracy is reasonable when compared with state-of-the-art work using similar experimental settings, i.e., tested on completely unseen areas [5]. More accurate LCZ classification results can be expected if larger amounts of high-quality reference data are available in the future, as the LCZ reference might still contain errors, even after manual editing by expert annotators. Future work includes exploring the potential of MTL for more test scenes and longer time series to provide support for applications, such as environmental management and monitoring.

V. Conclusion

HSE density and LCZ maps are both vital for urban analysis. Based on the intuition that both tasks are highly correlated and might provide hints for each other, we proposed to jointly predict HSE density and LCZ labels via an MTL approach and developed an MTL-based CNN to better leverage the complementary features extracted from the input satellite imagery. We validated the proposed approach with extensive experiments using Sentinel-2 data in the context of large-scale urban analysis. This letter shows that related urban mapping tasks can be performed jointly for improved generalization ability of DL models.

References

[1] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bull. Amer. Meteorol. Soc.*, vol. 93, no. 12, pp. 1879–1900, Dec. 2012.

[2] C. Corbane *et al.*, "Big Earth data analytics on Sentinel-1 and Landsat imagery in support to global human settlements mapping," *Big Earth Data*, vol. 1, nos. 1–2, pp. 118–144, Dec. 2017.

[3] T. Esch *et al.*, "Urban footprint processor—Fully automated processing chain generating settlement masks from global data of the TanDEM-X mission," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1617–1621, Nov. 2013.

[4] C. Zhang *et al.*, "Joint deep learning for land cover and land use classification," *Remote Sens. Environ.*, vol. 221, pp. 173–187, Feb. 2019.

[5] M. Demuzere, B. Bechtel, and G. Mills, "Global transferability of local climate zone models," *Urban Climate*, vol. 27, pp. 46–63, Mar. 2019.

[6] B. Bechtel *et al.*, "Towards consistent mapping of urban structures-global human settlement layer and local climate zones," in *Proc. 23rd ISPRS Congr.*, 2016, pp. 1371–1378.

[7] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*. [Online]. Available: http://arxiv.org/abs/1706.05098

[8] L. Liebel, K. Bittner, and M. Körner, "A generalized multi-task learning approach to stereo DSM filtering in urban areas," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 213–227, Aug. 2020.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[11] S. Kohl *et al.*, "A probabilistic U-Net for segmentation of ambiguous images," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 6965–6975.

[12] C. Qiu, M. Schmitt, C. Geiß, T.-H.-K. Chen, and X. X. Zhu, "A framework for large-scale mapping of human settlement extent from Sentinel-2 images via fully convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 152–170, May 2020.

[13] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491.

[14] M. Guo, A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei, "Dynamic task prioritization for multitask learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 282–299.

[15] T. Langanke, "Copernicus land monitoring service–high resolution layer imperviousness: Product specifications document," in *Proc. Copernicus EEA*, 2016, p. 12.

[16] *Data Source of the HSE Reference*. Accessed: Jan. 20, 2019. [Online]. Available: https://land.copernicus.eu/pan-european/high-resolution-layers/imperviousness

[17] *LCZ Reference From WUDAPT*. Accessed: Mar. 30, 2020. [Online]. Available: http://www.wudapt.org/continental-lcz-maps/

[18] X. X. Zhu *et al.*, "So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [software and data Sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 3, pp. 76–89, Sep. 2020.