Deep Semi-Supervised Multi-Task Learning of Building Features for District Energy Demand Estimation

Haozhen Cheng*
Karlsruhe Institute of Technology, Institute for Automation and Applied Informatics
Karlsruhe, Germany
haozhen.cheng@kit.edu

Jan Hoffmann
Karlsruhe Institute of Technology, Institute for Automation and Applied Informatics
Karlsruhe, Germany
jan.hoffmann@kit.edu

Hüseyin K. Çakmak
Karlsruhe Institute of Technology, Institute for Automation and Applied Informatics
Karlsruhe, Germany
hueseyin.cakmak@kit.edu

Veit Hagenmeyer
Karlsruhe Institute of Technology, Institute for Automation and Applied Informatics
Karlsruhe, Germany
veit.hagenmeyer@kit.edu

Abstract

Building energy demand is a primary driver of greenhouse gas emissions, necessitating accurate, sector-coupled energy system analysis via co-simulation. However, parameterizing white-box building models is frequently hindered by the unavailability of essential features in public databases. We present a method to estimate missing attributes—specifically construction year (CY), building type (BT), and energy carrier (EC)—using street-level imagery (SLI) across Germany as an exemplary use case. Our automated workflow integrates SLI with 2022 German Census data for labeling and is validated against OpenStreetMap (OSM) building geometries. A single deep learning model is developed using multi-task learning (MTL) and combined with semi-supervised learning (SSL) to effectively leverage partially labeled datasets. While the results demonstrate strong generalization to unseen test data, performance remains constrained by data quality issues that impact full automation.

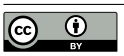
CCS Concepts

• **Computing methodologies** → **Object recognition**; • **Information systems** → **Information retrieval**.

Keywords

Semi-Supervised Learning, Multi-Task Learning, Deep Learning, Census Data, Street-level image data, Open data, Building models

ACM Reference Format:
Haozhen Cheng, Jan Hoffmann, Hüseyin K. Çakmak, and Veit Hagenmeyer. 2026. Deep Semi-Supervised Multi-Task Learning of Building Features for District Energy Demand Estimation. In *The 17th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '26), June 22–25, 2026, Banff, AB, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3744255.3811730



E-Energy '26, Banff, AB, Canada

ACM ISBN 979-8-4007-2011-6/26/06
https://doi.org/10.1145/3744255.3811730

1 Introduction

Germany's residential sector accounts for over 25% of final energy consumption and significant greenhouse gas emissions [42], making its management crucial for achieving national net-zero neutrality by 2045 [43]. Simulating city district energy systems enables the evaluation of heating alternatives in terms of environmental and economic feasibility [10]. Through co-simulation, integrated models can analyze sector-coupled electrical and heat grids [8] using bottom-up, white-box, multi-physics approaches that parameterize generic building envelopes and equipment [11]. However, real-world modeling often lacks the required building-specific data, necessitating the estimation of these features from alternative sources.

1.1 Contribution

This study advances the application of integrating street-level imagery (SLI) into urban energy system modeling through several key scientific and methodological contributions:

- **Automated Large-Scale Data Integration:** We propose a robust, automated workflow for the acquisition and labeling of SLI by combining synthetic images with high-resolution census data.
- **Synergistic Deep Learning Architecture:** We introduce a multi-task learning framework based on SSL enhancement, specifically optimized for parameterization bottleneck in white-box building models for sector-coupled co-simulation.
- **Expansion of Feature Estimation Scope:** We demonstrate the simultaneous estimation of three critical heating system attributes (CY, BT, and EC), providing a more comprehensive dataset for large-scale energy demand modeling.

The new methodology is presented using Germany as an example and can be applied to other countries with sufficient input data.

1.2 Related Work

The estimation of CY and BT from SLI has emerged as a prominent field, driven by the deployment of Convolutional Neural Networks (CNN) architectures such as ResNet [1, 14, 34, 44, 48], DenseNet

[24, 40, 41], and EfficientNet/V2 [21, 47]. Furthermore, advancements in multi-task learning [14] and Transformer-based architectures, notably the Swin Transformer [35], have significantly enhanced the granularity of building feature analysis. Nevertheless, achieving high predictive accuracy remains a persistent challenge. In particular, due to inherent variability in the dataset and limited geographic coverage, CY estimation is often limited to the range of 60%–70% [35, 48]. While BT estimation is less extensively explored, higher accuracy levels—approximating 85%—have been reported in localized studies [34]. A bottleneck exists regarding the spatial scalability of these models. Research by Benz et al. [1] highlights a significant degradation in model performance when transitioning from local applications to nationwide scales—a phenomenon also observed by Huang et al. [21]. To date, no framework has attempted to scale SLI-based building feature estimation to a nationwide scale, representing a major research gap that this study aims to fill.

2 Methods

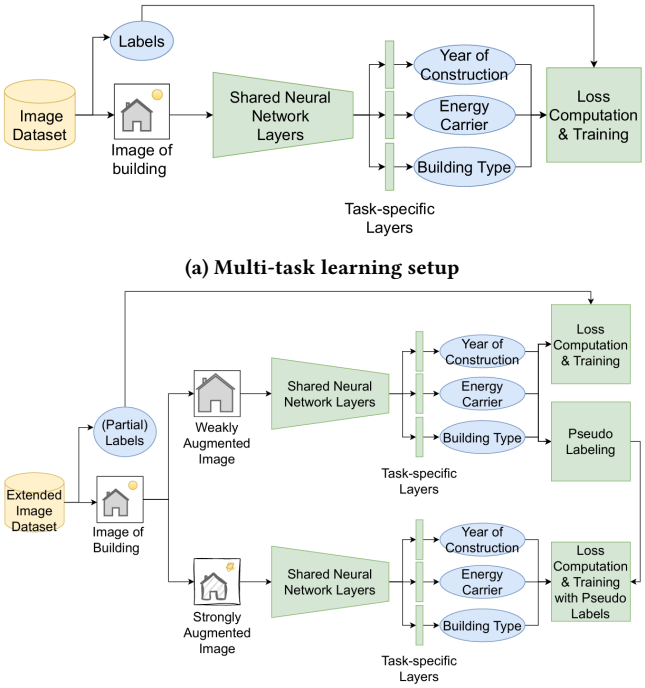
2.1 Workflow for Building Features

An automated workflow integrates 2022 German Census data[1] with Mapillary street-level imagery[2] to construct a labeled image dataset (Section 2.2). OSM[3] building geometry is utilized to validate feature visibility within the retrieved images. This dataset trains a multi-task deep learning model (Section 2.3) using SSL to accommodate incomplete labeling in census records. The model simultaneously estimates CY, BT, and EC for building heating systems. Specifically, the estimated CY and BT drive a U-value lookup via the TABULA project [30] to define the heat transfer coefficient. These parameters, combined with OSM-derived features (e.g., footprint and stories), enable the parameterization of white-box building models for sector-coupled co-simulation of urban electrical and heat grids.

2.2 Dataset Construction Workflow

The dataset construction follows a four-stage workflow. First, building feature data is extracted from the 2022 German Census to establish spatial coordinates. Second, street-level imagery is retrieved corresponding to these census locations. Third, OSM footprint data is used for geometric validation to ensure that buildings are visible within the image ranges. Finally, a quality-control filter is applied to remove low-quality images, yielding the final labeled dataset.

2.2.1 Census Data Extraction. Building features are sourced from the 2022 German census [4] with a 100m resolution grid covering Germany. We focus on CY (decades) [6] and BT [5]. To facilitate classification, feature values are mapped to integers; these mappings are detailed in Appendix A.1 for CY (Table 1), BT (Table 2), and EC (Table 3). In this process, infrequent or low-precision features are merged into larger classes, while unrepresentable values are omitted. To establish a reliable label source, we extend the "homogeneous cell" concept used in [2, 13, 16, 18, 45]. We define *fully homogeneous cells* as those reporting a single value across all three features, and *partly homogeneous cells* as those reporting a single value for only one or two features. Although privacy-preserving



(a) Multi-task learning setup

(b) Semi-Supervised multi-task learning setup

Figure 1: Visual overview of our deep learning methods

measures may obscure true homogeneity by zeroing out specific counts, the census "cell-key method" [37] ensures that columns derived from identical underlying data entries are transformed in the same way. Consequently, in a truly homogeneous cell, the count in a feature-specific column must match the aggregate building count for that cell. We select only those cells satisfying this condition as reliable indicators of homogeneity.

2.2.2 Image Data Retrieval. While commercial services like Google Street View[4] or Apple Look Around[5] are unsuitable due to restrictive terms of service [22, 29], Mapillary and Kartaview[6] provide user-contributed imagery under the permissive CC-BY-SA license [25, 31]. We select Mapillary as our primary source due to its higher availability across Germany. We query the Mapillary application programming interface (API) using a 150m × 150m bounding box—centered on the census cell with an additional 25m buffer in each direction. Beyond raw imagery, the API provides semantic segmentation masks; we use building detections within these masks to crop source images into "cutouts" containing only building pixels. This process may produce multiple cutouts from a single source image. To filter out small image cutouts with only a few pixels, the ratio of the number of pixels in a detection to the number of pixels in the entire image is computed. The threshold is set to 0.7%.

[1] https://www.destatis.de/EN
[2] https://www.mapillary.com/
[3] https://www.openstreetmap.org
[4] https://www.google.com/intl/en/streetview/
[5] https://maps.apple.com/imagecollection
[6] https://kartaview.org

2.2.3 Validation of Building Visibility with OpenStreetMap Data. We implement a validation workflow adapted from [9], integrating Mapillary API camera coordinates and compass angles with building footprints retrieved via the Overpass API. To resolve the $360°/0°$ discontinuity inherent in simple bearing methods [9], we identify the largest angular gap between bearings to define the building's angular coverage. Unlike the intersection tests used in [9], our approach determines visibility based on camera distance: closer buildings are prioritized, while distant buildings that are partially or fully occluded are clipped or removed. Buildings are subsequently filtered by census cell membership and residential usage [4], with generic `building=yes` tags treated as residential to mitigate data loss. Final validation compares these angular arcs with building-detection positions (left, center, or right) from Section 2.2.2. An image is accepted only if a "desired" building is detected within a $50°$ threshold and no "undesired" building occupies the same range. A wider $100°$ threshold checks for invisibility of undesired buildings to compensate for missing Mapillary FOV metadata.

2.2.4 Improving Data Quality. Manual analysis of the retrieved image data reveals a wide range of quality issues. In addition to general quality issues such as blurriness, distortion, or extreme brightness/darkness, we also find issues related to ambiguity in the depicted buildings, insufficient facade information, and the depiction of non-building objects. Our approach to automatically detecting images with these quality issues can be viewed as identifying different classes of outliers. We therefore employ the unsupervised isolation forest (IF) method [26] for outlier detection. Our approach is similar to [32], using a CNN as a feature extractor and then constructing an IF from each image's feature vectors. The IF then produces an anomaly score for each feature vector and, consequently, for each image. We examine different thresholds for outlier classification in Section 3.2. Images classified as outliers are then blacklisted from the dataset.

2.3 Deep Learning Methods

This section introduces the methods for training a deep learning model using the dataset created with the methods presented in Section 2.2. For the implementation, we use the PyTorch library[7].

2.3.1 Neural Network Architecture. We examine four neural network architectures as a backbone for our deep learning model: ResNet50 [19], DenseNet161 [20], ConvNeXt Tiny [28] and Swin Transformer Tiny [27] with 23.5M, 26.5M, 27.8M and 27.5M parameters respectively. We employ implementations and pre-trained weights obtained from training with the ImageNet dataset [15] from the torchvision library[8].

2.3.2 Multi-Task Learning. We use multi-task learning to estimate all three building features CY, BT, and EC with a single neural network. As noted previously, we estimate the CY, BT, and EC of the heating system to enable more robust multi-task learning. The fully connected classification layer used for ImageNet classification, present in all four networks, is replaced with a separate fully connected layer for each task. All other layers are shared across the tasks, so the classification layers receive the same feature vector as input (see Figure 1a).

2.3.3 Semi-Supervised Learning. To leverage imagery from partially homogeneous cells—where certain tasks lack labels—we implement a semi-supervised multi-task adaptation of FixMatch [39]. Following [33], we treat all images uniformly via a sampling approach that ignores the distinction between labeled and unlabeled data. Our framework extends the FixMatch mechanism to multi-task learning [36]: supervised loss and task-specific pseudo-labels are computed from weakly augmented images and then used to regularize predictions on strongly augmented inputs, yielding a second loss component. Both loss components are then summed to train the neural network. For strong augmentations, we employ RandAugment [12], adapting the implementation from [23]. The complete semi-supervised setup is illustrated in Figure 1b.

2.3.4 Hyperparameter Optimization. Using the Optuna library[9], we perform automated hyperparameter tuning for all four models with a fixed 24-hour compute budget. The complete configurations and optimal values are detailed in Appendix A.2. While most parameters are automated, the number of epochs is manually fixed at 30 for fine-tuning. Batch sizes are set to 128 for all architectures except the Swin Tiny-based model, which utilizes a batch size of 256 to enhance training stability.

3 Evaluation

3.1 Experimental Setup

Three datasets are constructed following Section 2.2. The *baseline* (178 860 images; 70-15-15 split for training, validation, and testing) comprises fully homogeneous cells. The *quality-improved* set applies the refinement from Section 2.2.4 to the baseline test subset (9 905 images). The *extended* dataset augments this with 300 377 images of partially homogeneous cells, totaling 426 927 training samples. To prevent leakage, all datasets are split at the census cell level; to improve evaluation quality, splitting preceded refinement to ensure comparability. All experiments use an NVIDIA A100 (40 GB) GPU, with performance measured via accuracy, F1-score, and confusion matrices.

3.2 Performance with Improved Data Quality

To assess image quality, an IF is applied to 900 randomly sampled images (100 per age class), revealing that 43.3% exhibited at least one quality issue. The resulting precision-recall curve (Figure 2, Appendix A.5) yields an average precision of 0.68; while effective at identifying anomalies, the method also flags several high-quality images. Based on this, three recall-driven contamination thresholds (0.5, 0.7, and 0.8) are evaluated—corresponding to the exclusion of 33%, 52%, and 64% of the most anomalous dataset—to generate the *quality-improved dataset (qid)*. ResNet-50 models, utilizing baseline hyperparameters (Appendix A.3), are trained on these subsets and evaluated as presented in Table 10 (Appendix A.6). Evaluating on the *qid* test set significantly improves performance over the baseline, by about 18% for CY and 4% for BT (row 1 vs. row 2). However, training on increasingly filtered datasets (rows 3-5) leads to a

[7] https://pytorch.org/
[8] https://docs.pytorch.org/vision/main/models.html
[9] https://optuna.org/

performance decline (e.g. for high contamination *(qid)* the reduction in Accuracy:−6.07%, F1:−7.26% for CY and Accuracy:−4.38%, F1:−3.02% for BT compared to baseline model *(qid)*, likely due to the reduced sample volume caused by higher contamination thresholds. Consequently, we use the full dataset for training and the *qid* test set for all subsequent evaluations.

3.3 Neural Network Architecture Comparison

Architectures from Section 2.3.1 are evaluated using optimized hyperparameters (Section 2.3.4; Appendix A.4) on the *qid* test set. As shown in Table 11 in Appendix A.7, all optimized models outperform the baseline model with Accuracy:0.448, F1:0.424 for CY and Accuracy:0.875, F1:0.837 for BT estimation. ResNet50 with optimized hyperparameters achieves the highest accuracy and F1-score for both tasks (CY Accuracy:0.465, F1:0.441 and BT Accuracy:0.882, F1:0.849) and is therefore selected for all subsequent evaluations.

3.4 Semi-Supervised Learning

To evaluate the impact of partially labeled data from the *extended dataset*, a FixMatch-based semi-supervised method (Section 2.3.3) is implemented using the ResNet-50 architecture and optimized hyperparameters (Section 3.3). Results, obtained on the *quality-improved dataset* test set, are presented in Table 12 (Appendix A.8). Compared to the fully supervised baseline (Table 11), the semi-supervised approach yields a significant performance increase for the CY task. In contrast, BT performance remained largely unchanged. This difference stems from the greater availability of CY labels in the extended dataset (Section 3.1), which may indicate saturation in BT classification performance.

3.4.1 Confusion Matrix: Construction Year and Building Type. Confusion matrices for the optimal semi-supervised model are provided in Appendix A.10, following the mapping in Appendix A.1. For the CY task, two primary trends emerge: first, misclassifications frequently occur between proximal age classes, suggesting potential utility if categories are broadened; second, systematic errors in classes 6 and 7 likely stem from class imbalance due to their lower sample density. Regarding the BT task, a notable misclassification of class 1 as class 0 is observed. Interestingly, despite the higher frequency of class 1, error rates for both classes remain comparable, suggesting that class imbalance is not the primary driver of this specific error.

3.4.2 Limitations of Energy Carrier Estimation. EC estimation yields substandard results due to two primary factors: the inherent semantic gap between building appearance and heating system components, and significant class imbalance. While gas accounts for 53.9% of German heating systems [7], our dataset overrepresents district heating. This is a sampling artifact resulting from our use of homogeneous census cells; since district heating networks cover entire neighborhoods, they generate more spatially homogeneous cells than other ECs. Using the ResNet50 model trained on the *extended dataset* (Section 3.4), results are presented in Table 13 (Appendix A.9). Although accuracy appears moderate, the sharp divergence from the F1-score indicates poor predictive reliability, a trend confirmed by the confusion matrix in Figure 5 (Appendix A.10). While classes 0 (Gas) and 5 (District Heating) show predictive capability, class 1 (Oil) exhibits higher error rates than correct classifications. For all other ECs, the model performs significantly worse, frequently defaulting to class 0 predictions. Consequently, robust EC estimation remains unfeasible with the current approach.

3.5 Scalability and Computational Efficiency

A case study for the city of Wuppertal, Germany (55 600 residential buildings, 364 776 residents as of 2025) highlights the workflow scalability. From 6 779 Mapillary API queries, 321 020 potential images are retrieved, yielding 262 442 cutouts after filtering. Although the API failure rate reached 16.7%, our automated error-detection algorithm seamlessly mitigates these instances. While image retrieval is the primary bottleneck (43h 55m), OSM-based validation is highly efficient (7m 31s), refining the dataset to 10 160 verified images. Using the pre-trained ResNet-50/FixMatch model (Section 3.4), inference on this new dataset takes only 18 seconds, demonstrating high throughput for large-scale urban deployment. The model training requires 10h 20m on an *AI-HPC cluster*.

4 Discussion

4.1 Comparison to Literature

Our peak CY accuracy (51%) is lower than the 60-70% range reported in Section 1.2; however, dataset heterogeneity prevents direct comparison. This stems from three factors: (i) greater spatial variation due to our nationwide study scope compared to locally limited studies at the city level; (ii) higher task complexity arising from narrow-band, single-decade classes rather than broader groupings; and (iii) the inherent data-quality challenges detailed in Section 4.3.

4.2 Data Quality Challenges

We identify three dimensions of data quality challenges: source uncertainty, methodological constraints, and structural ambiguity.

First, source-level characteristics introduce significant variance. The volunteer-contributed nature of Mapillary results in high stochasticity regarding lighting and sensor quality. Additionally, OSM tagging errors—specifically missing residential tags—lead to the inclusion of non-residential structures. While census cell selection poses risks of both masked intra-cell heterogeneity (Section 2.2.1) and geographic bias [18], our semi-supervised framework partially mitigates these risks by incorporating diverse, unlabeled imagery.

Second, methodological limitations affect the efficacy of quality control. Our current detection approach is suboptimal, as it incurs high precision costs by discarding high-quality samples and suffers from low recall under high contamination thresholds.

Third, using Mapillary segmentation masks for computational efficiency can introduce structural ambiguity (Section 2.2.4), since multiple buildings per detection can introduce errors in highly heterogeneous environments. Another source of uncertainty is that the input data may not be temporally aligned.

4.3 Deep Learning Architectures

ResNet-based models outperform more sophisticated architectures, such as Swin Transformer and ConvNeXt. This disparity is attributed to constraints on hyperparameter optimization; simpler

architectures allow for more efficient tuning within our fixed computational budget. These findings suggest that, for this specific task, increased architectural complexity does not yield proportional performance benefits. Future research should explore models that exceed our 30M-parameter limit—specifically, foundation models such as DINOv3 [38]—and evaluate the impact of alternative MTL and SSL configurations [17, 46].

5 Conclusion and Outlook

This work presents an automated workflow for estimating building features at a nationwide scale. Our deep learning model demonstrates robust generalization to unseen territories, with performance further optimized through hyperparameter tuning and SSL. However, data quality remains a critical bottleneck: high-fidelity, clean inputs are paramount for reliable results. Future research should prioritize advanced quality detection techniques and the integration of multiple data sources. Furthermore, scaling model architectures and extending these methods to the building parameterization and co-simulation workflows introduced in Section 1 represents a significant opportunity for further development.

Acknowledgments

This work is conducted within the framework of the Helmholtz Program Energy System Design (ESD) in the project "Helmholtz platform for the design of robust energy systems and their supply chains" (RESUR) and supported by the Helmholtz Association's Initiative and Networking Fund on the HAICORE@KIT and HAICORE@FZJ partitions.

References

[1] Alexander Benz, Conrad Voelker, Sven Daubert, and Volker Rodehorst. 2023. Towards an automated image-based estimation of building age as input for Building Energy Modeling (BEM). *Energy and Buildings* 292 (Aug. 2023), 113166. doi:10.1016/j.enbuild.2023.113166

[2] Luis Blanco, Megha Aditya, Björn Schiricke, and Bernhard Hoffschmidt. 2023. Classification of building properties from the German census data for energy analyses purposes. In *Proceedings of Building Simulation 2023: 18th Conference of IBPSA (Building Simulation, Vol. 18)*. IBPSA, Shanghai, China, 817–824. doi:10.26868/25222708.2023.1266

[3] Statistisches Bundesamt. 2024. Ergebnisse des Zensus 2022 - Gebäude mit Wohnraum nach Energieträger der Heizung. https://www.destatis.de/static/DE/zensus/gitterdaten/Zensus2022_Energietraeger.zip

[4] Statistisches Bundesamt. 2025. Datensatzbeschreibung zu den Tabellen "Gebäude nach Baujahr (Jahrzehnte) in Gitterzellen". https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Zensus2022/Publikationen/Downloads-Publikationen/Gitterdaten/datensatzbeschreibung_gebaeude_baujahr_jahrzente.xlsx?__blob=publicationFile&v=2

[5] Statistisches Bundesamt. 2025. Ergebnisse des Zensus 2022 - Gebäude mit Wohnraum nach Gebäudetyp (Größe). https://www.destatis.de/static/DE/zensus/gitterdaten/Gebaeude_mit_Wohnraum_nach_Gebaeudetyp_Groesse.zip

[6] Statistisches Bundesamt. 2025. Ergebnisse des Zensus 2022 - Gebäude nach Baujahr (Jahrzehnte). https://www.destatis.de/static/DE/zensus/gitterdaten/Gebaeude_nach_Baujahr_Jahrzehnte.zip

[7] Statistisches Bundesamt. 2025. Gebäude: Energieträger der Heizung. https://ergebnisse.zensus2022.de/datenbank/online/statistic/3000G/table/3000G-1008

[8] Hüseyin Kemâl Çakmak, Alexander Kocher, Haozhen Cheng, Jovana Kovačević, and Veit Hagenmeyer. 2025. *Collaborative Modelling and Co-Simulation for Sector-Coupled Multi-Energy System Analysis with eASiMOV-eCoSim*. doi:10.5445/IR/1000182805 37.12.02; LK 01.

[9] N. Çelik and E. Sümer. 2020. GEO-TAGGED IMAGE RETRIEVAL FROM MAPILLARY STREET IMAGES FOR A TARGET BUILDING. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIV-4/W3-2020 (2020), 151–158. doi:10.5194/isprs-archives-XLIV-4-W3-2020-151-2020

[10] Haozhen Cheng, Verena Buccoliero, Alexander Kocher, Veit Hagenmeyer, and Hüseyin K. Çakmak. 2024. New Co-Simulation Variants for Emissions and Cost Reduction of Sustainable District Heating Planning. In *2024 IEEE PES 16th Asia-Pacific Power and Energy Engineering Conference (APPEEC)*. 1–5. doi:10.1109/APPEEC61255.2024.10922291

[11] Haozhen Cheng, Jan Stock, André Xhonneux, Hüseyin K. Çakmak, and Veit Hagenmeyer. 2025. Construction and Control of Validated Highly Configurable Multi-Physics Building Models for the Sustainability Analysis of Multi-Energy Systems in a Co-Simulation Setup. In *SoutheastCon 2025*. 19–28. doi:10.1109/SoutheastCon56624.2025.10971478

[12] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. 2020. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 18613–18624. https://proceedings.neurips.cc/paper_files/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf

[13] Kristina Dabrock, Jens Ulken, Noah Pflugradt, Jann Michael Weinand, and Detlef Stolten. 2025. Generating a nationwide residential building types dataset using machine learning. *Building and Environment* 274 (April 2025), 112782. doi:10.1016/j.buildenv.2025.112782

[14] Menglin Dai, Jakub Jurczyk, Hadi Arbabi, Ruichang Mao, Wil Ward, Martin Mayfield, Gang Liu, and Danielle Densley Tingley. 2024. Component-Level Residential Building Material Stock Characterization Using Computer Vision Techniques. *Environmental Science & Technology* (Feb. 2024), acs.est.3c09207. doi:10.1021/acs.est.3c09207

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. doi:10.1109/CVPR.2009.5206848

[16] Ariane Droin, Michael Wurm, and Wolfgang Sulzer. 2020. Semantic labelling of building types A comparison of two approaches using Random Forest and Deep Learning. In *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e.V.*, Thomas P. Kersten (Ed.). Deutsche Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e.V., Stuttgart, Deutschland, 527–538. https://www.dgpf.de/src/tagung/jt2020/proceedings/proceedings/papers/76_KKNP_DGPF2020_Droin_et_al.pdf ISSN: ISSN 0942-2870.

[17] Maxime Fontana, Michael Spratling, and Miaojing Shi. 2024. When Multitask Learning Meets Partial Supervision: A Computer Vision Review. *Proc. IEEE* 112, 6 (2024), 516–543. doi:10.1109/JPROC.2024.3435012

[18] Oana M. Garbasevschi, Jacob Estevam Schmiedt, Trivik Verma, Iulia Lefter, Willem K. Korthals Altes, Ariane Droin, Björn Schiricke, and Michael Wurm. 2021. Spatial factors influencing building age prediction and implications for urban residential energy modelling. *Computers, Environment and Urban Systems* 88 (July 2021), 101637. doi:10.1016/j.compenvurbsys.2021.101637

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. doi:10.1109/CVPR.2016.90

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269. doi:10.1109/CVPR.2017.243

[21] Weimin Huang, Alexander W. Olson, Elias B. Khalil, and Shoshanna Saxe. 2025. Image-based prediction of residential building attributes with deep learning. *Journal of Industrial Ecology* 29, 1 (Feb. 2025), 81–95. doi:10.1111/jiec.13591

[22] Apple Inc. 2024. Apple Maps Terms of Use. https://www.apple.com/legal/internet-services/maps/terms-en.html

[23] Jungdae Kim. 2020. PyTorch implementation of FixMatch. https://github.com/kekmodel/FixMatch-pytorch

[24] Yan Li, Yiqun Chen, Abbas Rajabifard, Kourosh Khoshelham, and Mitko Aleksandrov. 2018. Estimating Building Age from Google Street View Images Using Deep Learning. In *10th International Conference on Geographic Information Science (GIScience 2018) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 114)*, Stephan Winter, Amy Griffin, and Monika Sester (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 40:1–40:7. doi:10.4230/LIPIcs.GISCIENCE.2018.40

[25] Meta Platforms Ireland Limited. 2025. Mapillary Terms of Use. https://www.mapillary.com/terms

[26] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*. 413–422. doi:10.1109/ICDM.2008.17

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9992–10002. doi:10.1109/ICCV48922.2021.00986

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11966–11976. doi:10.1109/CVPR52688.2022.01167

[29] Google LLC. 2025. Google Maps Platform Terms of Service. https://cloud.google.com/maps-platform/terms

[30] Tobias Loga, Nikolaus Diefenbach, Britta Stein, and Rolf Born. 2012. TABULA - Scientific Report Germany - Further Development of the National Residential Building Typology. https://episcope.eu/fileadmin/tabula/public/docs/scientific/DE_TABULA_ScientificReport_IWU.pdf Publisher: Institut Wohnen und Umwelt GmbH.

[31] Grabtaxi Holdings Pte. Ltd. 2025. Terms of Use for KartaView. https://kartaview.org/terms

[32] Siyu Luan, Zonghua Gu, Leonid B. Freidovich, Lili Jiang, and Qingling Zhao. 2021. Out-of-Distribution Detection for Deep Neural Networks With Isolation Forest and Local Outlier Factor. *IEEE Access* 9 (2021), 132980–132989. doi:10.1109/ACCESS.2021.3108451

[33] Miquel Marti, Sebastian Bujwid, Alessandro Pieropan, Hossein Azizpour, and Atsuto Maki. 2022. An analysis of over-sampling labeled data in semi-supervised learning with FixMatch. *Proceedings of the Northern Lights Deep Learning Workshop* 3 (April 2022). doi:10.7557/18.6269 Publisher: UiT The Arctic University of Norway.

[34] Ryan Murdoch and Ala'a Al-Habashna. 2024. Residential building type classification from street-view imagery with convolutional neural networks. *Signal, Image and Video Processing* 18, 2 (March 2024), 1949–1958. doi:10.1007/s11760-023-02882-8

[35] Yoshiki Ogawa, Chenbo Zhao, Takuya Oki, Shenglong Chen, and Yoshihide Sekimoto. 2023. Deep Learning Approach for Classifying the Built Year and Structure of Individual Buildings by Automatically Linking Street View Images and GIS Building Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), 1740–1755. doi:10.1109/JSTARS.2023.3237509

[36] Lei Qi, Hongpeng Yang, Yinghuan Shi, and Xin Geng. 2024. MultiMatch: Multi-task Learning for Semi-supervised Domain Generalization. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 6, Article 184 (March 2024), 21 pages. doi:10.1145/3648680

[37] Stefanie Setzer, Johannes Rohde, Volker Güttgemanns, and Patrick Rothe. 2024. Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. - Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens. *WISTA - Wirtschaft und Statistik* (June 2024). https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2024/03/cell-key-methode-teil1-032024.html

[38] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. 2025. DINOv3. arXiv:2508.10104 [cs.CV] https://arxiv.org/abs/2508.10104

[39] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. FixMatch: simplifying semi-supervised learning with consistency and confidence. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.). Curran Associates Inc., Red Hook, NY, USA, Article 51, 13 pages. https://proceedings.neurips.cc/paper_files/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf

[40] Maoran Sun, Fan Zhang, and Fabio Duarte. 2021. Automatic Building Age Prediction from Street View Images. In *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*. IEEE, Beijing, China, 102–106. doi:10.1109/IC-NIDC54101.2021.9660554

[41] Maoran Sun, Fan Zhang, Fabio Duarte, and Carlo Ratti. 2022. Understanding architecture age and style through deep learning. *Cities* 128 (Sept. 2022), 103787. doi:10.1016/j.cities.2022.103787

[42] Umweltbundesamt. 2025. Energieverbrauch privater Haushalte. https://www.umweltbundesamt.de/daten/private-haushalte-konsum/wohnen/energieverbrauch-privater-haushalte

[43] Umweltbundesamt. 2025. Treibhausgasminderungsziele Deutschlands. https://www.umweltbundesamt.de/daten/klima/treibhausgasminderungsziele-deutschlands

[44] W.O.C. Ward, X. Li, Y. Sun, M. Dai, H. Arbabi, D. Densley Tingley, and M. Mayfield. 2023. Estimating energy consumption of residential buildings at scale with drive-by image capture. *Building and Environment* 234 (April 2023), 110188. doi:10.1016/j.buildenv.2023.110188

[45] Michael Wurm, Ariane Droin, Thomas Stark, Christian Geiß, Wolfgang Sulzer, and Hannes Taubenböck. 2021. Deep Learning-Based Generation of Building Stock Data from Remote Sensing for Urban Heat Demand Modeling. *ISPRS International Journal of Geo-Information* 10, 1 (Jan. 2021), 23. doi:10.3390/ijgi10010023

[46] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. 2023. A Survey on Deep Semi-Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2023), 8934–8954. doi:10.1109/TKDE.2022.3220219

[47] Xinran Yu, Zhengbo Zou, and Semiha Ergan. 2023. Extracting principal building variables from automatically collected urban scale façade images for energy conservation through deep transfer learning. *Applied Energy* 344 (Aug. 2023), 121228. doi:10.1016/j.apenergy.2023.121228

[48] Matthias Zeppelzauer, Miroslav Despotovic, Muntaha Sakeena, David Koch, and Mario Döller. 2018. Automatic Prediction of Building Age from Photographs. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval* (Yokohama, Japan) *(ICMR '18)*. Association for Computing Machinery, New York, NY, USA, 126–134. doi:10.1145/3206025.3206060

A Appendix

A.1 Mapping of Census Feature Values to Classes

In the following, the mapping of feature values from the census data to the classes we use for classification can be found for all three building features in Table 1, Table 2, and Table 3, respectively.

Census Feature Value	Class Mapping
< 1919	0
1919 - 1949	1
1950 - 1959	2
1960 - 1969	3
1970 - 1979	4
1980 - 1989	5
1990 - 1999	6
2000 - 2009	7
2010 - 2015	8
≥ 2016	8

Table 1: Left: Feature values reported in the census data for construction year [6], Right: Mapping to Integer Classes

Census Feature Value	Class Mapping
Detached Single-Family Home	0
Semi-Detached Single-Family Home	0
Terraced Single-Family Home	0
Detached Double-Family Home	1
Semi-Detached Double-Family Home	1
Terraced Double-Family Home	1
Multi-Family Home (3-6 Apartments)	1
Multi-Family Home (7-12 Apartments)	1
Multi-Family Home (≥13 Apartments)	1
Different Building Type	

Table 2: Left: Feature values reported in the census data for building type [5], Right: Mapping to Integer Classes

Census Feature Value	Class Mapping
Gas	0
Oil	1
Wood & Wood Pellets	2
Biomass & Biogas	
Solar & Geothermics & Heat Pumps	3
Electric	4
Coal	
District Heating	5
No Enery Carrier	

Table 3: Left: Feature values reported in the census data for energy carrier [3], Right: Mapping to Integer Classes

A.2 List of Hyperparameters

We optimize the hyperparameters seen in Table 4 separately for all four models. In this table, the hyperparameters we considered can be found in the left column. In the right column are all possible choices for each of these hyperparameters. Most choices include the definition of further hyperparameters which are listed below them.

Hyperparameter	Values
Optimizer	**SGD** learning rate: $\eta \in [10^{-4}, 1.0]$ weight decay: $\lambda \in [10^{-4}, 1.0]$ momentum: $\beta \in [0.0, 0.99]$ **AdamW** learning rate: $\eta \in [10^{-5}, 0.01]$ weight decay: $\lambda \in [10^{-4}, 1.0]$ first-order moment: $\beta_1 \in [0.85, 0.95]$ second-order moment: $\beta_2 \in [0.98, 0.9999]$
Learning Rate Scheduler	**Use Learning Rate Scheduler** warmup epochs $T_{\text{init}} \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ initial learning rate $\eta_{\text{init}} \in [10^{-6}, \eta]$ minimal learning rate ratio $r_{\eta_{\min}} \in [10^{-3}, 0.1]$
Loss Function	**Cross Entropy Loss** **Symmetric Cross Entropy Loss** CE Weight: $\alpha \in [0.01, 1.0]$ RCE Weight $\beta \in [0.1, 5.0]$
Model Layer Freeze	**Freeze Number of Layers** $\in \{0, 1, 2, 3, 4, 5, 6\}$

Table 4: List of Hyperparameters considered for optimization

A.3 Hyperparameters of Baseline Model

The hyperparameters we use for the baseline model can be found in Table 5. We use a fixed learning rate in this case, therefore the learning rate scheduler is missing in the table.

Hyperparameter	Values
Optimizer	**AdamW** learning rate = 0.001 weight decay = 0.0001 beta1 = 0.9 beta2 = 0.999
Loss Function	**Cross Entropy Loss**
Model Layer Freeze	**Freeze Number of Layers** = 0

Table 5: Baseline Hyperparameters

A.4 Optimized Hyperparameters

The optimized hyperparameters for the different models can be found in the following tables. These hyperparameters are used without rounding when training the models for the evaluation. Thus, in order to ensure reproducibility, the unrounded hyperparameters are given here as well. For hyperparameters where there is a choice between multiple options, the option that has been found to perform best is shown here. Other options are thus removed from the table. If the entry for the learning rate scheduler is missing in the table, this means that it is not used and a fixed learning rate is selected.

Hyperparameter	Values
Optimizer	**AdamW** learning rate = 0.0013701784704900643 weight decay = 0.018824819022270234 beta1 = 0.9187153188059649 beta2 = 0.986930824776534
Learning Rate Scheduler	**Use Learning Rate Scheduler** number of warmup epochs = 10 initial learning rate = 0.0006611157164129546 minimal learning rate ratio = 0.008630891525973686
Loss Function	**Symmetric Cross Entropy Loss** alpha = 0.04299971789690542 beta = 0.11911829532343003
Model Layer Freeze	**Freeze Number of Layers** = 4

Table 6: Optimized Hyperparameters for ResNet50

Hyperparameter	Values
Optimizer	**AdamW** learning rate = 0.0002476820180532099 weight decay = 0.7209117337718214 beta1 = 0.876218016738078 beta2 = 0.9979453974224378
Learning Rate Scheduler	**Use Learning Rate Scheduler** number of warmup epochs = 4 initial learning rate = $3.1039928937017984 \cdot 10^{-5}$ minimal learning rate ratio = 0.0034885267568203755
Loss Function	**Symmetric Cross Entropy Loss** alpha = 0.4301465662429676 beta = 0.291611574644313
Model Layer Freeze	**Freeze Number of Layers** = 0

Table 7: Optimized Hyperparameters for Swin Tiny

Hyperparameter	Values
Optimizer	**AdamW** learning rate = 0.00017408255301165225 weight decay = 0.0018418282374452059 beta1 = 0.9181528921664687 beta2 = 0.983241542737244
Loss Function	**Symmetric Cross Entropy Loss** alpha = 0.4100939534863066 beta = 1.9904884203892608
Model Layer Freeze	**Freeze Number of Layers** = 3

Table 8: Optimized Hyperparameters for ConvNeXt Tiny

Hyperparameter	Values
Optimizer	**AdamW** learning rate = 0.00010431087032652898 weight decay = 0.23366267192570556 beta1 = 0.8501340167044523 beta2 = 0.9802660039428862
Loss Function	**Cross Entropy Loss**
Model Layer Freeze	**Freeze Number of Layers** = 6

Table 9: Optimized Hyperparameters for DenseNet161

A.5 Precision-Recall Curve for Anomaly Detection using the Isolation-Forest Method

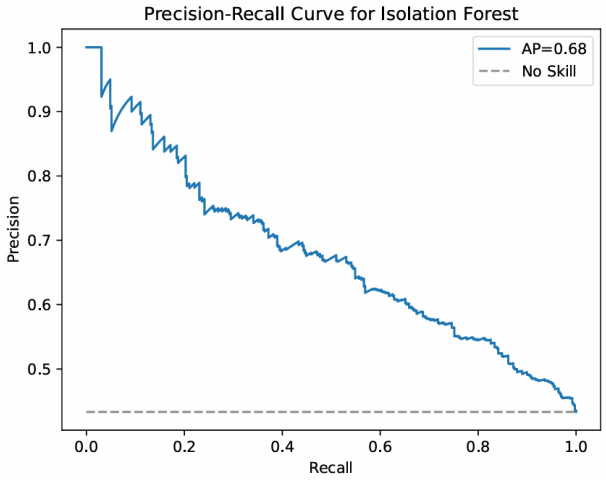


Figure 2: Precision-Recall curve for detecting anomalous images with the isolation forest-based method

A.6 Accuracy and F1-Scores with Improved Data Quality

	Construction Year		Building Type	
	Accuracy	F1-Score	Accuracy	F1-Score
Baseline model	0.3811	0.3570	0.8373	0.8078
Baseline model *(qid)*	0.4480	0.4244	0.8752	0.8374
Low contamination *(qid)*	0.4407	0.4098	0.8787	0.8451
Medium contamination *(qid)*	0.4247	0.3802	0.8707	0.8369
High contamination *(qid)*	0.4208	0.3936	0.8576	0.8121

Table 10: Accuracy and F1-Scores comparing different contamination levels used in the training process. Higher is better. First row: Testing subset of the baseline dataset, remaining rows: Testing subset of the *quality-improved dataset (qid)*

A.7 Accuracy and F1-Scores for Architectures

	Construction Year		Building Type	
	Accuracy	F1-Score	Accuracy	F1-Score
Baseline model (ResNet50) *(qid)*	0.4480	0.4244	0.8752	0.8374
Optimized ResNet50 *(qid)*	0.4659	0.4414	0.8820	0.8491
Optimized Swin Tiny *(qid)*	0.4539	0.4271	0.8726	0.8393
Optimized ConvNeXt Tiny *(qid)*	0.4397	0.4045	0.8776	0.8459
Optimized DenseNet161 *(qid)*	0.4630	0.4392	0.8768	0.8388

Table 11: Accuracy and F1-Scores for different neural network architectures and optimized hyperparameters. Higher is better. All rows use the testing subset of the *quality-improved dataset (qid)*

A.8 Accuracy and F1-Scores for CY and BT

	Construction Year		Building Type	
	Accuracy	F1-Score	Accuracy	F1-Score
Optimized ResNet50 *(qid)*	0.4659	0.4414	0.8820	0.8491
Optimized ResNet50 + SSL *(qid)*	0.5101	0.4813	0.8810	0.8499

Table 12: Accuracy and F1-Scores for training without and with semi-supervised learning. All rows use the testing subset of the *quality-improved dataset (qid)*

A.9 Accuracy and F1-Scores for EC

	Energy Carrier	
	Accuracy	F1-Score
Optimized ResNet50 + SSL	0.6232	0.3305

Table 13: Accuracy and F1-Scores for the energy carrier task

A.10 Confusion Matrices for SSL

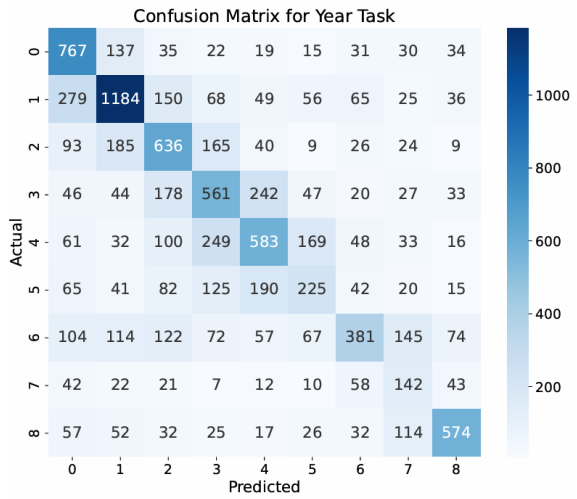


Figure 3: Semi-Supervised Learning (SSL): Confusion matrix for construction year task. Accuracy: 51.01%

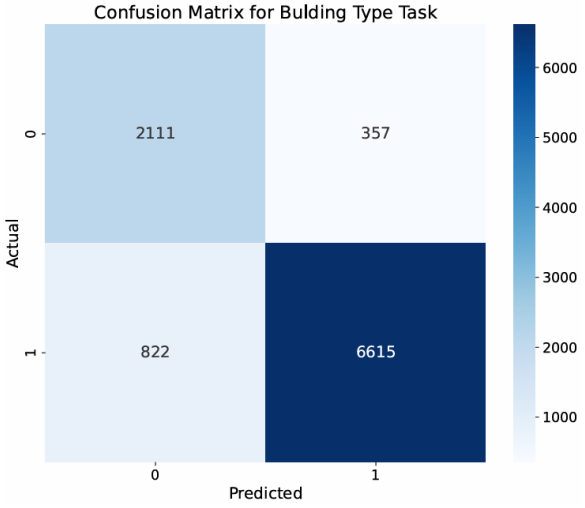


Figure 4: Semi-Supervised Learning: Confusion matrix for building type task. Accuracy: 88.10%

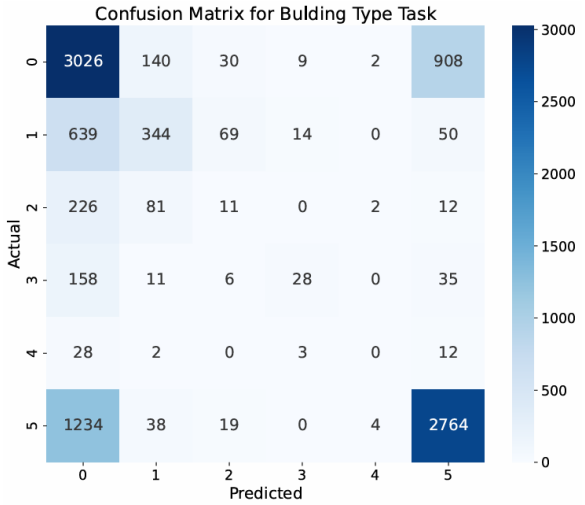


Figure 5: Semi-Supervised Learning: Confusion matrix for energy carrier task. Accuracy: 62.32%