



EG-ICE 2025 GLASGOW

Decoding Building Facades Automated Element and Material Recognition in Street-Level Images for Large-Scale Building Stock Assessments

LEONHARD NOLL^{1,2}, FLORIAN NOICHL^{1,2}, BEYZA KIPER³, ANDRÉ BORRMANN^{1,2}

¹Chair of Computing in Civil and Building Engineering, Technical University of Munich, Germany

²TUM Georg Nemetschek Institute, Technical University of Munich, Germany

³NYU Tandon School of Engineering, New York University, USA

ABSTRACT

As a basis for environmental assessments like Life Cycle Assessment (LCA) of large building portfolios, extensive image data of building envelopes must be evaluated as automatically as possible. This paper addresses the automated detection of both elements – e.g., windows, walls, doors – and materials in building facades with unified machine learning workflows using 2D RGB images. Following a systematic review of existing methods and datasets, two unified segmentation workflows are developed: Hierarchical Segmentation (HS) and Multi-Task Learning (MTL). HS exploits the hierarchical relationships between facade elements and materials and deploys a post-prediction clustering approach with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), while MTL leverages shared feature learning for simultaneous detection. To mitigate limited training data, this work introduces the high-resolution segmentation and classification dataset *Facades Material Munich* (FaMatMuc). For the first time, element and material detection for facade images were combined in one workflow and validated successfully.

KEYWORDS

Facade, Computer vision, Semantic segmentation, Building materials, Life Cycle Assessment

1. INTRODUCTION

The construction sector contributes significantly to greenhouse gas (GHG) emissions, and therefore offers considerable potential for effecting meaningful change, especially given the aging building stock. Evaluating the energy performance of existing buildings and developing effective retrofitting strategies are imperative to achieve substantial reductions in emissions and energy costs. Life Cycle Assessment (LCA) is the standard methodology for the assessment of environmental impacts and is gaining prominence due to the EU's sustainability reporting

requirements for larger companies. However, the prevailing practice of manual extraction of building geometry and characteristics from architectural plans and on-site inspections remains a time-consuming process, impeding scalability for the assessment of larger building stocks (Dai *et al.*, 2021). Previous research reveals a significant gap: the absence of element-material segmentation workflows, alongside the lack of high-resolution, diverse datasets for material detection. This paper aims to address this research gap by introducing an automated approach, leveraging a unified workflow for

detecting elements – e.g., windows, walls, doors – and materials in building facades, thereby enhancing efficiency and scalability. Automated extraction of data on facade elements and materials enables downstream processes by enriching simple building representations, thereby facilitating the population of the inventory analysis phase for LCAs and providing essential inputs for energy simulation tools that model diverse retrofit interventions. Additionally, we provide an openly available dataset that enables future research in the field and is used to demonstrate the efficacy of the developed methods.

2. RELATED WORKS

Automating the analysis of building facades has become a major area of research, driven by the increasing demand for efficient and scalable methods to support sustainability efforts. Recent advancements in deep learning and the integration of diverse data sources have enabled more comprehensive methods for automated assessment of existing buildings. For example, Forth, Noichl and Borrmann (2024) proposed a multimodal approach that integrates point clouds and RGB data with a component database to create an enriched surface model with data relevant for LCA. Such semantic enrichment of raw input data requires the identification of building elements like windows, doors, roofs, and others, as well as the materials to obtain the necessary information for environmental assessments.

2.1 Computer vision for facade analysis

Detecting building elements, in conjunction with facade parsing (semantic segmentation of facade images) has shown notable advancements.

Modern deep learning techniques, such as Convolutional Neural Networks (CNNs) and, more recently, Vision Transformers (ViTs), have surpassed traditional rule-based approaches for facade parsing. In this, it has been shown that ViTs for semantic segmentation of building elements achieve superior performance on established benchmarks in comparison to all previous methods. (Wang *et al.*, 2024)

Despite these advancements, research addressing the semantic segmentation of materials in building facades remains limited. In a previous study, Habibi *et al.* (2022) introduced a dataset comprising close-up images and hyperspectral data of facades from an industrial area in Australia; however, there is currently no publicly available semantic segmentation dataset for facade materials that encompasses a diverse range of architectural styles and is suitable for full-facade analysis.

A promising approach for material facade parsing was proposed by Xu *et al.* (2023). Since distinguishing features between materials relies more on textures and patterns than on shapes and colors, the

authors addressed this using a Multi-Scale Contextual Attention Network, which incorporated transformer attention mechanisms to exploit details from different scales. However, the approach has several limitations. First, the reliance on the simplified assumption that each building consists of only two primary materials and the challenges of annotating hybrid facades led to classification inconsistencies. Second, the low image resolution of 2046 x 2046 pixels and the depiction of multiple buildings in a single frame is inadequate for the extraction of fine textures and patterns, a main objective of the authors' methodology. As a result, the generated segmentation masks lack sufficient granularity, which makes this approach unsuitable for applications such as LCA that require finer detail.

Raghu, Bucher and De Wolf (2023) investigated material classification of entire facades by conducting a comparative analysis of three state-of-the-art neural network architectures: a transformer-based model, a hybrid CNN-transformer architecture, and a purely CNN-based model. These models were evaluated on a newly created classification dataset with images from five cities. The dataset consists of non-rectified images sourced from Google Street View, which were annotated image-wise with one or more classes. None of the applied models consistently outperformed the others on all data sets. Furthermore, the study was limited to image-level classification and did not include any form of segmentation.

This overview of related works on facade element and material recognition reveals that no previous approach has combined these dual challenges using a unified workflow or facilitated shared information across tasks.

2.2 Datasets

The stated approaches employ supervised learning methods. To enable such applications, there is a need for annotated datasets for the training and evaluation of segmentation methods. It is important to note that the content and labeling approach in such a dataset must match the task at hand.

Table 1 offers a comprehensive overview of the main characteristics of available datasets for detecting elements or materials in building facades. It includes annotated datasets that are publicly accessible or made accessible for this research. Several annotated datasets for building elements have been published, whereas the two material datasets indicate that material detection in building facades has received comparatively little attention. All datasets vary significantly in terms of observed class types, annotation quality, resolution, and the way facades are captured (single buildings vs. multiple buildings, rectified vs. oblique picture frame).

Table 1: Overview of publicly available facade element and material datasets with key characteristics.

Refs.: [1] Teboul et al. (2010), [2] Korč and Förstner (2009), [3] Tylecek and Sára (2013), [4] Frohlich, Rodner and Denzler (2010), [5] Sun et al. (2022), [6] Kong and Fan (2021), [7] Wang et al. (2024), [8] Gadde, Marlet and Paragios (2016), [9] Habili et al. (2022), [10] Raghu, Bucher and De Wolf (2023). SeSe = semantic segmentation, InSe= instance segmentation, OA = object annotation. Common class types abbreviated as s: sky, w: window, d: door

Dataset	Size	Annotation		Resolution	Image Properties		Details
		Type	Class types		Single Building	Rectified	
ECP [1]	104	SeSe	7: { <u>s</u> , chimney, roof, <u>w</u> , balcony, wall, <u>d</u> , shop}	404 x 640	✓	✓	Hausmannian architecture
eTRIMS [2]	60	SeSe, InSe	8: { <u>w</u> , wall, <u>d</u> , <u>s</u> , pavement, vegetation, car, road}	768 x 512 512 x 768	✓	✓	Diverse architecture, mainly residential in Germany, Switzerland
CMP [3]	606	SeSe, OA	11: {facade, molding, cornice, pillar, <u>w</u> , <u>d</u> , sill, blind, balcony, shop, deco}	variable low	✗	✗	Diverse architecture, mostly stone and plaster/mortar facades
LabelMe-Facade [4]	945	SeSe	8: {building, car, <u>d</u> , pavement, road, <u>s</u> , vegetation, <u>w</u> }	683 x 512	✗	✗	Diverse architecture, poor annotation quality
Deep-Windows [5]	1200	SeSe	1: { <u>w</u> }	variable	✗	✗	Concatenated dataset
FacadeWHU [6]	900	SeSe, OA	6: { <u>w</u> , <u>d</u> , wall, balcony, road, shop}	variable	✗	✗	Diverse architecture, strong fisheye distortion
CFP [7]	602	InSe	9: {building, <u>w</u> , <u>d</u> , roof, tree, <u>s</u> , people, car, sign}	variable high	✗	✗	Diverse architecture
Paris Art Deco Facades [8]	79	SeSe	7: { <u>d</u> , shop, balcony, <u>w</u> , wall, <u>s</u> , roof}	variable low	✓	✓	Art-deco buildings in Paris
LIB HSI [9]	513	SeSe	9: {miscellaneous, vegetation, glass, <u>w</u> , brick, concrete, blocks, metal, <u>d</u> , timber}	512 x 512	✗	✗	Light industrial bldgs., close-up, good annotation quality; addn. hyperspectral data
Urban Resource Cadas-ter [10]	972	Multi-Label Classif.	8: {brick, stucco, rustication, siding, wood, metal, null, other}	640 x 400	✗	✗	Diverse architecture from NYC, Tokyo, Zurich

3. METHODOLOGY

This paper addresses the research gaps in detecting elements and materials in building facades with limited training data, by establishing two unified workflows handling both tasks and introducing a novel dataset. The contributions of this research are as follows, reflecting the order in which they were developed and implemented:

- Creation of a unified building element dataset: Three popular building element datasets ECP, CFP, eTRIMS were merged using a unified taxonomy, minimizing the need for additional annotations and allowing simple integration of heterogeneous element datasets for training.
- Introduction of the FaMatMuc Dataset (Facades Materials Munich): A new dataset was created, comprising 100 annotated high-resolution images of building facades in Munich, Germany for material semantic segmentation and 541 extracted patches for material classification. The dataset and documentation are available at <https://github.com/fnoi/famatmuc>.

- Development of unified semantic segmentation approaches for element and material detection: Two semantic segmentation methodologies, namely the Hierarchical Segmentation and a Multi-Task Learning method, were developed, allowing for sequential and simultaneous prediction of elements and materials in building facades.

3.1 Hierarchical Segmentation

The proposed Hierarchical Segmentation (HS) method is a sequential method that builds upon the hierarchical relationships between the various elements and materials that comprise a building facade. Unlike conventional approaches that rely heavily on color and shape, this method employs patch-based textural and pattern analysis for detection, thereby fully exploiting the high detail of the developed FaMatMuc dataset.

The methodology is visualized in Fig. 1. In the first step, instance segmentation for element detection is performed with Mask2Former (Cheng et al.,

2022) and the identified facade mask is isolated, removing openings and surroundings that do not contain any information about the facade material. Next, an optimization algorithm extracts patches by minimizing non-informative (black) pixels, ensuring that each patch captures unique and relevant facade regions with a maximum threshold of 10% black pixels. In the case of FaMatMuc, these patches have a size of 256 x 256 pixels to generally fit within the spaces between windows, as these wall sections are usually the narrowest ones in a facade.

Extracted patches are processed in Swin Transformer V2 (Liu *et al.*, 2022) for the initial material classification. Each patch is stored with its corresponding predicted label and top-left corner coordinates. This mixed type data, containing numerical coordinates and categorical predictions, is combined and clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello, Moulavi and Sander, 2013) to determine arbitrary-shaped clusters while considering the hierarchical structure and different densities within each cluster.

The Gower distance metric is used to homogenize the mixed data (Gower, 1971). This metric measures the dissimilarity between data points with mixed numerical and categorical variables. For this, input data is pre-processed including ordinal encoding of categorical labels and the normalization of spatial coordinates. The Gower metric is defined as:

$$D_{\text{Gower}}(x, y) = 1 - \frac{1}{m} \sum_{j=1}^m s_j(x, y), \quad (1)$$

where m represents the total number of variables, and $s_j(x, y)$ is a similarity function for the j -th variable, which is defined based on the type of variable:

$$s_j = \begin{cases} 1, & \text{if } j \text{ is categorical and } x_j = y_j, \\ 0, & \text{if } j \text{ is categorical and } x_j \neq y_j, \\ 1 - \frac{|x_j - y_j|}{R_j}, & \text{if } j \text{ is numerical.} \end{cases} \quad (2)$$

R_j denotes the range of the numerical variable j . Gower's distance is bounded within the interval [0,1], where 0 indicates that two data points are identical, and 1 indicates maximum dissimilarity. The hyperparameters for the patch clustering algorithm were selected based on the recommendations of the developers of HDBSCAN. The minimum cluster size is set to be at least three patches or 5% of the total number of patches. The minimum samples parameter, which defines the number of neighboring points required for a data point to be

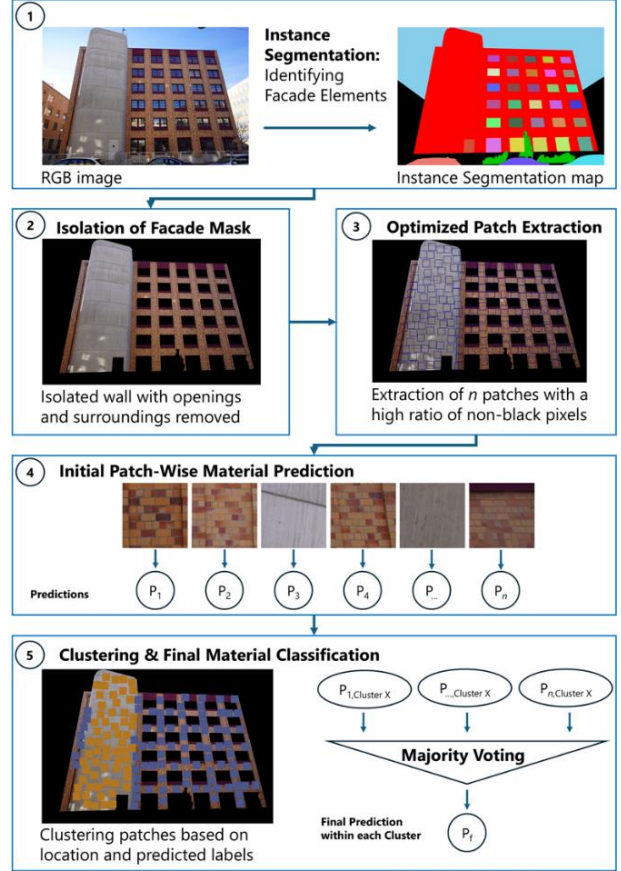


Fig. 1: HS workflow; from RGB input to material clusters via segmentation, patch extraction, clustering and majority-voting classification

considered a core point, is set to the maximum of 1 point or 5% of the total number of patches.

The HDBSCAN algorithm is applied to determine coherent patch clusters. Within each cluster, a majority voting mechanism is applied on the patches to assign the most frequently predicted label to the entire cluster. The result is a detailed map of the present facade materials, minimizing classification noise stemming from individual patches.

However, it is important to note that since the extracted patches do not cover the entire facade and are passed through a classifier, this method does not aim to perform pixel-wise material segmentation. Instead, it offers a practical approach to estimate the present materials, their location and coverage based on close-up facade patches, providing valuable information for further analysis.

3.2 Multi-Task Learning

The second presented approach particularly addresses the challenges of material detection and limited data availability by facilitating shared information across prediction tasks and offering simultaneous prediction of elements and materials in building facades. For implementing the Multi-Task

Learning (MTL) method, DeepLabv3+ (Chen *et al.*, 2018) was selected due to its competitive performance on urban street scene datasets and its adaptability to MTL approaches that allowed changes in the network’s architecture. To address the dual-task, a task-specific decoder is designed with a corresponding head – the element head and the material head – to generate predictions for their respective tasks. Figure 2 illustrates the structure of the MTL framework. During the forward pass, the input image is processed through the shared ResNet-101 backbone and the Atrous Spatial Pyramid Pooling (ASPP) module, extracting rich multi-scale features. These features are passed to the corresponding decoder, element head, and material head, which produce outputs specific to each task.

The training process integrates both element and material segmentation tasks in a mixed-batch approach. Each batch contains samples from both tasks, which are distinguished by a task index. That allows the model to learn simultaneously while element and material segmentation losses are computed independently. The model is further improved with auxiliary loss heads, which provide additional supervision at intermediate layers. According to Zhao *et al.* (2017) the auxiliary loss helps deep networks to stabilize and optimize the training process. The auxiliary outputs are weighted by an auxiliary loss factor of $\lambda_{aux} = 0.4$ and integrated into the loss function, which is optimized by minimizing the cross-entropy loss. The choice for the auxiliary loss factor follows the practice of Zhao *et al.* (2017). Yang *et al.* (2024) states that while auxiliary losses slightly increase training time, inference time is not increased.

Cross-entropy losses are defined for the element and material task and are combined into one final objective function to optimize both tasks, see Eq. 3.

$$\mathcal{L}_{total} = \mathcal{L}_{mat} + \mathcal{L}_{elem} + \lambda_{aux}(\mathcal{L}_{mat,aux} + \mathcal{L}_{elem,aux}), (3)$$

where $\mathcal{L}_{mat,aux}$ and $\mathcal{L}_{elem,aux}$ are the auxiliary loss components for material and element segmentation at intermediate layers, respectively.

The developed MTL network facilitates simultaneous training and inference, which allows for predicting both the semantic segmentation masks of building elements and materials. By making use of shared features across tasks, the MTL approach simplifies training and inference. The method does not aim to surpass the latest mean Intersection over Union (*mIoU*) and accuracy metrics achieved by transformer networks on building element datasets but offers a unified framework with competitive results to demonstrate the potential of MTL in the context of facade parsing.

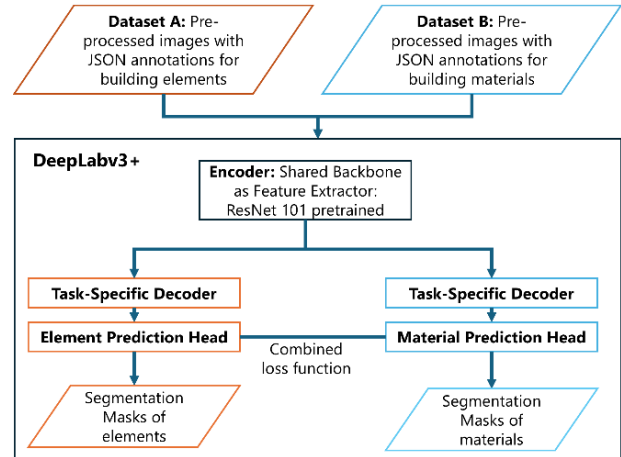


Fig. 2: Multi-task learning (MTL) semantic segmentation approach for building elements and materials using a modified DeepLabv3+ (Chen *et al.*, 2018) architecture: ResNet 101 encoder, task-specific prediction heads, and a combined loss function to optimize for both segmentation tasks

3.3 Development of FaMatMuc

The lack of annotated datasets for semantic segmentation and material detection in building facades is addressed through the development of a high-resolution dataset. To support the proposed approaches, a dedicated material dataset was created with two types of annotations, semantic segmentation masks and classification labels. The semantic segmentation dataset comprises 100 high-resolution images of building facades; the classification dataset includes 541 patches extracted from the captured images. This dataset primarily features single residential and office buildings taken from a frontal perspective in Munich, Germany. Efforts were made to ensure adequate representation of all targeted material classes (stucco, brick, stone, concrete, glass and timber). However, some class imbalance was observed due to the frequency of these materials in the city of Munich. An additional bias exists in the dataset, as the buildings captured for this dataset were selected based on accessibility and visual distinctiveness. The data was collected in January 2025 under clear, sunny weather conditions with mostly no leaves on trees and bushes.

The photos were taken using a Sony a7r III camera equipped with a Sony FE 16-35mm F/2.8 GM lens, enabling the capture of entire facades even from relatively short distances. Each image has a resolution of 5304 x 7952 pixels, which provides fine details for facade element segmentation and material classification. High-resolution images allow for the extraction of close-up areas of the captured photos while keeping a high level of detail, which supports the HS approach.



Fig. 3: Sample from the FaMatMuc dataset: The image has a resolution of 5304 x 7952 pixels and shows the front view perspective of a five-story residential building with a concrete base and a timber main facade

Due to the narrow streets and high buildings, most photos exhibit noticeable perspective distortion. Figure 3 shows a sample facade image from the collected dataset, Fig. 4 presents samples of the extracted patches. The images were used as-is for semantic segmentation labeling without prior undistortion. To maintain consistency in close-up patches for the hierarchical approach, 541 patches of size 256 x 256 were extracted from the high-resolution images prior to classification annotation. This patch size was chosen to align with the HS framework.

The proposed material classes were selected considering their prevalence in Central European architecture. Metal and plastic were excluded because they are less frequently observed and exhibit various appearances. These materials and any others not explicitly defined are included within the "other" class to ensure coverage of materials not represented in the predefined labels. Furthermore, a "background" class was introduced for semantic segmentation to segment parts of the image that do not provide material information about the facade. Table 2 presents a detailed description of the proposed classes.

To support the HS approach, 541 patches were annotated for the classification task; for the MTL approach, semantic masks were created to enable pixel-wise segmentation tasks. The annotation process was conducted by the first author and is validated by an additional expert.

The annotation strategy was designed to support LCA and energy performance analyses by allowing the identification and quantification of building materials significantly influencing environmental and energy-related outcomes. A special focus is laid on the necessary granularity, ensuring annotations are suitable for such applications.



Fig. 4: Samples of 10 representative patches (out of 541 extracted), each with a size of 256 x 256 pixels, annotated with their identified material labels

Thus, the primary goal is to facilitate the accurate processing of material areas in subsequent analysis steps while keeping the annotation process manageable in terms of time and effort. However, the calculation of the material area falls outside this research's scope. To achieve the necessary detail while being time-efficient, several key decisions were made to define the level of detail and streamline the workflow.

To remain practicable, minor functional components and occlusions such as rain gutters, street furniture, and vegetation were labeled according to the underlying material if identifiable. Assuming that omitting these elements would not significantly impact overall material area calculations but would instead improve surface area estimations. A streamlined approach was adopted for efficient annotation: if a facade wall consisted of a single material, the entire facade was selected as a single region, and inner elements, such as windows or doors, were "punched out" in a subsequent processing step. This was achieved by defining priority values for the materials when loading the data and drawing the semantic segmentation masks.

The following order is used: 1.) stucco 2.) brick 3.) stone 4.) concrete 5.) timber 6.) other 7.) glass. The time required to annotate a single image varied significantly depending on the complexity of the building facade but is estimated to be, on average,

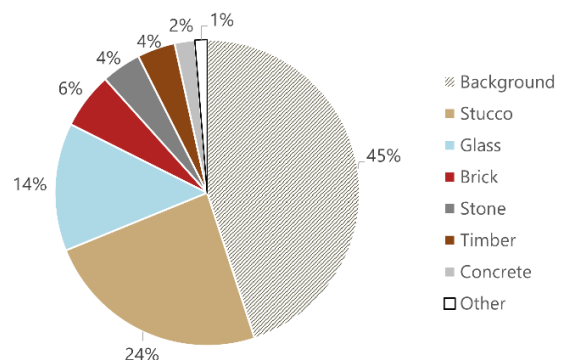


Fig. 5: Pixel-wise class distribution of FaMatMuc, containing 100 annotated images

approximately 15 minutes per image. Figure 5 shows the class distribution of the FaMatMuc

dataset for semantic segmentation. The distribution shows a significant class imbalance.

Table 2: Visual description, typical use and material properties of FaMatMuc material classes

Material	Visual Description	Typical Use / Properties
Stucco	Mostly uniform color, smooth and flat surface, fine to medium grain. Features include color uniformity and smooth texture.	Commonly used as a finish for composite insulation systems; durable but prone to cracking over time.
Brick	Regular pattern, usually reddish or brownish, with visible mortar lines. Features include color patterns and regular brick alignment.	Widely used in both traditional and modern facades; durable, low-maintenance, and recyclable.
Stone	Varied texture, rough or smooth; visible layers; usually larger than bricks. Features include texture variations, color differences, and block size.	Often used for decorative or structural purposes; highly durable and weather-resistant, recyclable.
Concrete	Gray, smooth, or slightly rough with occasional imperfections. Features include general color uniformity and surface imperfections.	Common in construction for structural and aesthetic purposes; strong, durable, and versatile; limited recyclability.
Glass	Reflective surface, often smooth and uniform. Features include reflectivity and general smoothness.	Common in buildings for windows and facades; provides transparency and aesthetic appeal; high recyclability.
Timber	Grainy patterns, often brownish or beige tones. Features include color variations and large-scale patterns.	Used in wooden cladding or decorative facades; requires maintenance to prevent decay; high recyclability.

3.4. Merged Building Element Dataset

To address the challenges of limited data availability and geographical restrictions, this work merges three existing element datasets using a unified taxonomy, creating the Merged Building Element Dataset (MBED), as shown in Tab. 3. Classes that are not relevant for facade segmentation are disregarded in this projection. Additionally, roof annotations were incorporated into the eTRIMS dataset to ensure uniform labeling across all datasets. Each dataset is assigned to a unique identifier, which allows the data loader to apply the corresponding taxonomy mapping during data loading.

In this study, MBED is used for training of element detection in the HS pipeline, as it substantially increases training data. For the MTL pipeline, MBED is not used, as it mainly consists of CFP images, which primarily feature Chinese architecture – including materials beyond the scope of this research. Additionally, CFP images differ significantly from typical image frames, often depicting multiple facades in a single shot, whereas eTRIMS and ECP typically contain only one facade per image.

Table 3: Unified taxonomy mapping for CFP (602 images), eTRIMS (60 images) and ECP (104 images); for references see Table 1

Dataset	Original Class	Projected Class
CFP	Building	Facade
	Window	Window
	Door	Door
	Roof	Roof
	Tree	Vegetation
	Sky	Sky
	People	-
	Car	Car
	Sign	-
	eTRIMS	Window
Wall		Facade
Door		Door
Sky		Sky
Pavement		-
Vegetation		Vegetation
Car		Car
Road		-
ECP	Sky	Sky
	Chimney	-
	Roof	Roof
	Window	Window
	Balcony	Balcony
	Wall	Facade
	Door	Door
	Shop	-

4. EXPERIMENTS & RESULTS

To validate the proposed methods, the models were trained and tested on ECP, merged ECP & eTRIMS, and further evaluated with a dedicated Case Study dataset, which is recorded in the same way as FaMatMuc. For this purpose, 12 additional images were annotated with semantic segmentation

masks for building elements and materials. The element annotations adhere to the style of the ECP and eTRIMS dataset. Although the limited number of 12 images does not allow for comprehensive statistical conclusions, they serve as a first indicator for validating the methodology.

The evaluation is based on common pixel-wise metrics for semantic segmentation: Intersection over Union (*IoU*), mean *IoU* (*mIoU*), accuracy, precision and recall (Everingham *et al.*, 2010).

4.1. Implementation

The presented project was developed in Jupyter Notebooks with an NGC container image of PyTorch 2.5.0 (Paszke *et al.*, 2019) to optimize for NVIDIA GPU acceleration. The training process was executed on an NVIDIA Tesla V100 with a single GPU (16 GB memory) as well as an NVIDIA Ampere A100 with 20 GB multi-instance GPU.

The datasets were split into training, validation and test sets to allow proper model development, which includes hyperparameter optimization on the validation set and performance evaluation on the test set. For training Mask2Former on the building element datasets, a random split (70%, 15%, 15%) with a fixed random seed of 42 was applied to ensure reproducible behavior. The random split was chosen due to the relatively balanced distribution of these datasets (Wang *et al.*, 2024). For the training of the Swin Transformer V2, the FaMatMuc classification dataset was split into training (90%) and test (10%) set, as the implementation follows the approach of (Raghu, Bucher and De Wolf, 2023). The building material dataset FaMatMuc, which exhibits a significant class imbalance, is divided into training, validation, and test sets using a structured, greedy pixel-constrained optimization to balance the class distribution across all subsets.

For this research, geometric transformations and color space augmentation were applied to expand the training dataset.

4.2. Results Hierarchical Segmentation

For evaluating HS, a method to evaluate clustered patch predictions against the semantic segmentation ground truth mask is defined, as the patch covers a certain area of the original image. A patch prediction is correct if the assigned material matches at least 40% of the patch area in the ground truth. This threshold balances strictness and tolerance for segmentation noise at material boundaries.

Following the HS approach, in the first step, images are processed through Mask2Former to get their segmentation maps. The reported metrics on the segmentation masks are limited to the classes *facade* and *window*, which are given in Tab. 4.

Table 4: *IoU* and accuracy per element class (excluding background) for the Case Study dataset, all metrics reported in %

Class Label	<i>IoU</i>	Accuracy
Facade	85.21	88.70
Window	68.00	83.06

For classification for materials, the Swin Transformer V2 was trained on the FaMatMuc classification dataset (541 patches). The results in Tab. 5 show excellent performance in all metrics, suggesting that the classification task on FaMatMuc is a rather easy task for the Swin Transformer V2.

Table 5: Metrics obtained from FaMatMuc’s classification dataset trained on Swin Transformer V2

Metric	Value [%]
Accuracy	99.08
Precision	99.10
Recall	97.96

Subsequently, the isolated facade masks are processed through the developed post-prediction clustering algorithm with HDBSCAN. As shown in Tab. 6, the final patch classification after clustering achieves strong performance metrics, with a particularly high precision score of 94.52%. This indicates reliable material identification when the model makes positive predictions.

Table 6: Classification metrics for the HS approach on the Case Study dataset

Metric	Value [%]
Accuracy	82.72
Precision	94.52
Recall	86.77

Figure 6 shows one sample with the visualization of the finalized clusters after majority voting.



Fig. 6: Visualization of finalized clusters on a patch map after applying majority voting within each cluster to determine dominant labels; here: decorative facade with stucco (beige) and brick (red) parts

This sample demonstrates that the method successfully identified primary material regions (brick, stucco) in the complex decorative facade. However, the method struggles to precisely delineate the material borders, especially when materials are interwoven.

4.3. Results Multi-Task Learning

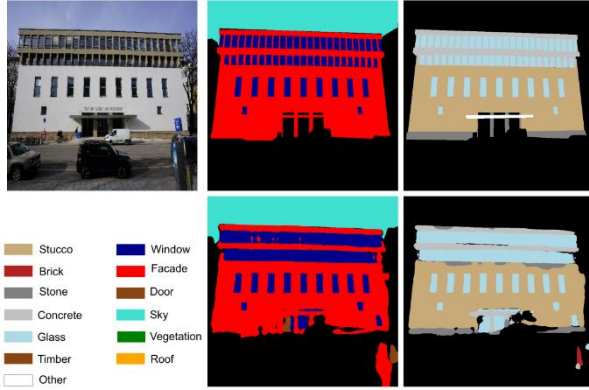


Fig. 7: Qualitative results of MTL-based semantic segmentation. Top row: (left) input image depicting the four-story main building of the Technical University Munich with a facade composed of stone, stucco, and concrete from bottom to top; (center) ground truth (GT) elements, (right): GT materials; Bottom row: (left) legend, (center) predicted element mask, (right) predicted material mask

Before evaluating MTL on the Case Study dataset, a Single-Task Learning (STL) approach is compared to our proposed MTL approach, as shown in Tab. 7. The STL network and its training setup are identical to the MTL approach, with the only difference being that the STL network only has one decoder for material prediction.

Table 7: Comparison of *mIoU*, mean accuracy, and overall accuracy for FaMatMuc: MTL vs. Single-Task Learning (STL), all metrics reported in %

Metric	MTL	STL	Δ
<i>mIoU</i>	64.51	63.28	+1.23
Mean Accuracy	76.86	72.49	+4.37
Overall Accuracy	89.36	88.97	+0.39

The MTL model achieves slightly better results on FaMatMuc with an improvement of 1.23% in *mIoU* compared to the STL network. This indicates that the material task benefits from the element task through shared learning. The proposed method shows the potential of shared feature learning in the context of facade parsing. In addition to performance improvements, the MTL model simplifies the training and inference by leveraging simultaneous learning and its adaptability to dual tasks.

Tables 8 and 9 show the calculated metrics achieved by the MTL-adapted DeepLabv3+ network for building elements and materials, respectively.

Table 8: *IoU* and accuracy per element class (excluding background) for the Case Study dataset, all metrics reported in %

Class Label	<i>IoU</i>	Accuracy
Facade	81.26	95.23
Window	70.87	78.80

Table 9: *IoU* and accuracy per class for FaMatMuc achieved with MTL, all metrics reported in %

Class Label	<i>IoU</i>	Accuracy
Stucco	80.72	90.06
Brick	83.62	90.21
Stone	36.35	60.35
Concrete	61.74	66.84
Glass	74.46	80.22
Timber	56.50	96.14

Stucco, glass, and brick, which are the three predominant classes in the Case Study, achieve the highest *IoU* scores. The relatively low *IoU* values of stone (36.35%), concrete (61.74%), and timber (56.50%) are not only due to their lower representation in the dataset but also likely caused by the visual similarity between stone and concrete. For qualitative inspection, one inference sample is given in Fig. 7.

4. CONCLUSION

This work successfully developed and validated two standalone semantic segmentation approaches for the simultaneous and sequential detection of elements and materials in building facades. To address the limited data availability, we harmonized three disparate building element datasets eTRIMS, ECP, and CFP using a unified taxonomy. As no material semantic segmentation dataset for the full facade analysis was available, this work developed the *Facades Material Munich* (FaMatMuc) semantic segmentation and classification dataset dedicated to the material detection of intra-urban facades of Central Europe. The high-resolution images offer a significantly superior resolution (5-10x higher) than the images in existing facade parsing datasets, which allow to identify patterns and textures for material detection.

A fundamental limitation of this work, common to most visual detection pipelines, results from the intrinsic limitations of RGB-based material detection. The absence of in-depth information due to the surface sensitivity of visual data limits the information that can be extracted. This underscores that visual detection should be complemented with additional contextual data to improve reliability and enhance

segmentation results. Despite the development of the high-resolution material segmentation dataset and the merge of three existing prominent element segmentation datasets, data scarcity and inconsistencies between element datasets remain an issue. However, its potential is limited by constraints for the model's input size imposed by the low-resolution nature of available building element datasets, as they cannot be drastically upscaled without quality loss.

ACKNOWLEDGEMENTS

This research has been partially funded by the GeoAI4Retrofit project supported by the Federal Ministry for Economic Affairs and Climate Action Germany.

REFERENCES

- Campello, R.J.G.B., Moulavi, D. and Sander, J. (2013) 'Density-Based Clustering Based on Hierarchical Density Estimates', in J. Pei et al. (eds) *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 160–172. DOI: https://doi.org/10.1007/978-3-642-37456-2_14.
- Chen, L.-C. et al. (2018) 'Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation', in V. Ferrari et al. (eds) *Computer Vision – ECCV 2018*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 833–851. DOI: https://doi.org/10.1007/978-3-030-01234-2_49.
- Cheng, B. et al. (2022) 'Masked-attention Mask Transformer for Universal Image Segmentation'. arXiv. Available: <http://arxiv.org/abs/2112.01527> [16 November 2024].
- Dai, M. et al. (2021) 'Residential building facade segmentation in the urban environment', *Building and Environment*, 199, p. 107921. DOI: <https://doi.org/10.1016/j.buildenv.2021.107921>.
- Everingham, M. et al. (2010) 'The Pascal Visual Object Classes (VOC) Challenge', *International Journal of Computer Vision*, 88(2), pp. 303–338. DOI: <https://doi.org/10.1007/s11263-009-0275-4>.
- Forth, K., Noichl, F. and Borrmann, A. (2024) 'LCA Calculation of Retrofitting Scenarios Using Geometric Model Reconstruction and Semantic Enrichment of Point Clouds and Images', in *Computing in Civil Engineering 2023. ASCE International Conference on Computing in Civil Engineering 2023*, Corvallis, Oregon: American Society of Civil Engineers, pp. 390–397. DOI: <https://doi.org/10.1061/9780784485231.047>.
- Frohlich, B., Rodner, E. and Denzler, J. (2010) 'A Fast Approach for Pixelwise Labeling of Facade Images', in *2010 20th International Conference on Pattern Recognition. 2010 20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey: IEEE, pp. 3029–3032. DOI: <https://doi.org/10.1109/ICPR.2010.742>.
- Gadde, R., Marlet, R. and Paragios, N. (2016) 'Learning Grammars for Architecture-Specific Facade Parsing', *International Journal of Computer Vision*, 117(3), pp. 290–316. DOI: <https://doi.org/10.1007/s11263-016-0887-4>.
- Gower, J.C. (1971) 'A General Coefficient of Similarity and Some of Its Properties', *Biometrics*, 27(4), p. 857. DOI: <https://doi.org/10.2307/2528823>.
- Habili, N. et al. (2022) 'LIB-HSI: RGB and Hyperspectral images of Building Facades'. CSIRO. DOI: <https://doi.org/10.25919/3541-S396>.
- Kong, G. and Fan, H. (2021) 'Enhanced Facade Parsing for Street-Level Images Using Convolutional Neural Networks', *IEEE Transactions on Geoscience and Remote Sensing*, 59(12), pp. 10519–10531. DOI: <https://doi.org/10.1109/TGRS.2020.3035878>.
- Korč, F. and Förstner, W. (2009) 'eTrims Image Database for Interpreting Images of Man-Made Scenes'. Available: http://www.ipb.uni-bonn.de/projects/etrims_db/.
- Liu, Z. et al. (2022) 'Swin Transformer V2: Scaling Up Capacity and Resolution'.
- Paszke, A. et al. (2019) 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. arXiv. Available: <https://doi.org/10.48550/arXiv.1912.01703>.
- Raghu, D., Bucher, M.J.J. and De Wolf, C. (2023) 'Towards a "resource cadastre" for a circular economy – Urban-scale building material detection using street view imagery and computer vision', *Resources, Conservation and Recycling*, 198, p. 107140. DOI: <https://doi.org/10.1016/j.resconrec.2023.107140>.
- Sun, Y. et al. (2022) 'DeepWindows: Windows Instance Segmentation through an Improved Mask R-CNN Using Spatial Attention and Relation Modules', *ISPRS International Journal of Geo-Information*, 11(3), p. 162. DOI: <https://doi.org/10.3390/ijgi11030162>.
- Teboul, O. et al. (2010) 'Segmentation of building facades using procedural shape priors', in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA: IEEE, pp. 3105–3112. DOI: <https://doi.org/10.1109/CVPR.2010.5540068>.
- Tylecek, R. and Sára, R. (2013) 'Spatial Pattern Templates for Recognition of Objects with Regular Structure'. Available: <https://cmp.felk.cvut.cz/~tylecr1/facade/>.
- Wang, B. et al. (2024) 'Improving facade parsing with vision transformers and line integration', *Advanced Engineering Informatics*, 60, p. 102463. DOI: <https://doi.org/10.1016/j.aei.2024.102463>.
- Xu, F. et al. (2023) 'Semantic segmentation of urban building surface materials using multi-scale contextual attention network', *ISPRS Journal of Photogrammetry and Remote Sensing*, 202, pp. 158–168. DOI: <https://doi.org/10.1016/j.isprsjprs.2023.06.001>.
- Yang, T. et al. (2024) 'LT-DeepLab: an improved DeepLabV3+ cross-scale segmentation algorithm for Zanthoxylum bungeanum Maxim leaf-trunk diseases in real-world environments', *Frontiers in Plant Science*, 15, p. 1423238. DOI: <https://doi.org/10.3389/fpls.2024.1423238>.
- Zhao, H. et al. (2017) 'Pyramid Scene Parsing Network', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, pp. 6230–6239. DOI: <https://doi.org/10.1109/CVPR.2017.660>.