ISPRS Journal of Photogrammetry and Remote Sensing 231 (2026) 357–375

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Bottom-up building exposure modeling with multimodal earth vision

Patrick Aravena Pelizari [a,*], Christian Geiß [a,b], Hannes Taubenböck [a,c]

[a] German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Weßling, 82234, Germany
[b] University of Bonn, Department of Geography, Bonn, 53115, Germany
[c] University of Würzburg, Institute of Geography and Geology, Department of Remote Sensing, Würzburg, 97074, Germany

ARTICLE INFO

Keywords:
Building exposure
Multimodal remote sensing
Street-level imagery
Multitask learning
Missing modality
Spatial context

ABSTRACT

Effective disaster mitigation and management rely on up-to-date exposure models providing detailed and spatially localized information on vulnerability-relevant characteristics of buildings. This study investigates the potential of heterogeneous multimodal geo-image data—incorporating street-level imagery (SLI), very high-resolution optical remote sensing data, and a normalized digital surface model—for generic large-area building characterization. We introduce a deep multimodal multitask learning methodology for the synergistic fusion of multi-sensor data and efficient multi-criteria building classification. The proposed *task-wise modality attention* (TMA) fusion optimizes multimodal feature representations for individual inference tasks separately according to their specific requirements. To address the challenge of partially missing SLI data (*i.e.*, the *missing modality problem*), a transformer-based *SLI spatial context encoder* leverages spatial correlations between structural building attributes and their visual manifestations to make the semantic information from available SLI widely accessible. With the earthquake-prone metropolis Santiago de Chile as test site, the two scenarios—*SLI available* and *SLI missing*—are evaluated through a comprehensive experimental cross-comparison of estimated generalization accuracies for classifying buildings according to five target variables: *height*, *lateral load-resisting system material*, *seismic building structural type*, *roof shape*, and *block position*. The results underscore the significant potential of the employed modalities and methods. Across the five addressed attributes, covering a total of 35 thematic classes, the most accurate models achieve mean κ accuracies of 85.19% and 74.96% for data points with and without SLI coverage, respectively. The presented data and methods allow to generate an area-wide building exposure model with a unique combination of thematic resolution, spatial detail and coverage.

1. Introduction

Population growth, urbanization, and climate change have led to a significant increase in the number of people and assets exposed to natural hazards worldwide (UNDRR, 2022; Dodman et al., 2022; Taubenböck et al., 2024). To understand, assess, and mitigate natural disaster risks, up-to-date and detailed knowledge of the exposed built environment—its spatial distribution and vulnerability—is essential (Wyss and Rosset, 2013). An exposure model includes a spatially referenced inventory of buildings, each assigned attributes defining its physical vulnerability to natural hazards (Taubenböck et al., 2009; Pittore et al., 2017). Alongside the information on the hazard itself, up-to-date exposure and vulnerability data are critical for designing adaptation strategies and disaster management plans before and after an event, based on risk analyses and damage assessments (Geiß and Taubenböck, 2013; UNISDR, 2015). However, due to the large number of buildings, their heterogeneous structural designs, and the spatio-temporal dynamics driven by urbanization, maintaining an inventory database across extensive areas is a highly complex task. Traditional data collection methods, such as in-situ building inspections, are not capable to meet this challenge (Pittore et al., 2017).

At the same time, holistic vulnerability assessments across multiple natural hazards impose high demands on exposure models in terms of thematic detail and spatial resolution, as *(i)* different building attributes may influence the vulnerability to different hazards (Silva et al., 2022), and *(ii)* natural hazards vary in spatial scale, exhibiting distinct spatial patterns and variabilities (Gill and Malamud, 2014; Dabbeek and Silva, 2019; Gómez Zapata et al., 2021). A generic description of the building stock combined with a high spatial resolution enhances the flexibility of the risk or impact model to consistently and efficiently address multi-hazard scenarios.

Building instances within an exposure model are assigned vulnerability-relevant characteristics according to standardized taxonomies (Pittore et al., 2018). *E.g.*, the GED4ALL multi-hazard building

* Corresponding author.
E-mail address: patrick.aravenapelizari@dlr.de (P. Aravena Pelizari).

https://doi.org/10.1016/j.isprsjprs.2025.10.036
Received 31 March 2025; Received in revised form 22 September 2025; Accepted 26 October 2025

classification system proposed by Silva et al. (2022) covers the *lateral load-resisting system* (LLRS; *i.e.*, the structural system that resists acting lateral forces such as seismic loads, wind loads, water pressure or earth pressure) and its material (*e.g.*, masonry or wood), height, occupancy, block position (*i.e.*, the position of a building or housing entity in relation to its neighbors), structural irregularity, and roof shape, among others. A vulnerability model (*e.g.*, a fragility curve) relates the intensity of a natural hazard to the damage probability of a building, as determined by its vulnerability-relevant characteristics. This enables the assessment of a building's vulnerability concerning a specific hazard intensity (Calvi et al., 2006; Douglas, 2007).

Drastic transformation processes, coupled with limited exposure data, require leveraging relevant datasets and developing automated methods to enable efficient vulnerability-related characterization of the built environment on a large scale. Driven by expanding data acquisition initiatives (both remote and in-situ sensing), social media, and advances in artificial intelligence, geospatial imaging sensor data has become a key source for automated spatial information extraction (Zhu et al., 2017; Ibrahim et al., 2020; Biljecki and Ito, 2021).

Numerous studies have demonstrated the potential of remote sensing data and supervised machine learning techniques for the spatially continuous extraction of vulnerability-relevant attributes at building object level using high to very high resolution sensors. Target variables include building height, occupancy, and roof type as well as the seismic building structural type (SBST), which characterizes a building's main load-bearing structure from a seismic vulnerability perspective (*e.g.*, Sarabandi and Kiremidjian, 2007; Geiß et al., 2015; Liuzzi et al., 2019; Zhou et al., 2023; Müller et al., 2023; Mutreja and Bittner, 2023; Li et al., 2024b,a).

By capturing the streetscape from a human vision perspective, street-level imagery (SLI) complements the top-down view of remote sensing data bridging information gaps that often hinder complex applications (Lefevre et al., 2017; Zhang et al., 2019; Biljecki and Ito, 2021)—*e.g.*, by providing high-resolution façade views. In their pioneering study, Wieland et al. (2012) employ omnidirectional SLI to extract structural attributes through expert-based visual image interpretation and to estimate building height via photogrammetric 3D building reconstruction. Compared to traditional in-situ surveys, such SLI-based remote visual screenings enable a decentralized and location-independent inspection of a large number of buildings, significantly increasing data collection efficiency (Geiß et al., 2017; Esquivel-Salas et al., 2022). This is particularly true when building upon commercial web mapping services (*e.g.*, Google Street View; Anguelov et al., 2010; Santa María et al., 2017; Pittore et al., 2018), crowd-sourcing based alternatives (*e.g.*, Mapillary or Kartaview; Hou et al., 2024) or social media (*e.g.*, Flickr; Hoffmann et al., 2023). The extraction of vulnerability-relevant building attributes using SLI and deep learning (DL) classification methods has been the subject of several studies, covering the identification of building height, LLRS, LLRS material, SBST, ductility, building age, roof shape, block position and soft-storey construction (Kang et al., 2018; Gonzalez et al., 2020; Yu et al., 2020; Qiao and Yuan, 2021; Aravena Pelizari et al., 2021, 2023; Sun et al., 2022; Ogawa et al., 2023), among others. Generalization accuracies show that combining geospatial imagery with machine learning-based inference enables efficient automated building inventory collection, providing a cost- and labor-efficient alternative to traditional methods. Also accounting for façade information, the derivation of structural building features from oblique images captured by unmanned aerial vehicles has been successfully demonstrated for spatially limited areas (Meng et al., 2021; Zhang et al., 2023). In the context of multi-hazard risk assessments, Aravena Pelizari et al. (2023) employ multitask learning to effectively address the need for generic building characterization at the algorithmic level, enabling the simultaneous prediction of multiple structural target variables with substantially enhanced model accuracy and efficiency.

However, a systematic study exploring the potential of integrating heterogeneous multimodal geo-image data for a generic, spatially continuous vulnerability-related characterization of buildings exposed to natural hazards remains absent. This paper aims to address this gap by considering SLI, very high-resolution (VHR) optical remote sensing data, and a normalized digital surface model (nDSM) derived from high-resolution optical imagery.

1.1. Multimodal geospatial imagery and deep learning

With the increasing availability of geospatial image data from different platforms and sensor types, along with derived products like digital surface models, the synergistic fusion of complementary modalities to enhance the target information has been a widely researched field (Gomez-Chova et al., 2015; Schmitt and Zhu, 2016; Aravena Pelizari et al., 2018; Hong et al., 2021; Li et al., 2022; Mena et al., 2024).

This study focuses on *heterogeneous data fusion*, *i.e.*, the integration of data derived from fundamentally different imaging mechanisms (*e.g.*, the fusion of SAR and optical data or the combination of remote sensing and ground-based data; Li et al., 2022). In this context, end-to-end optimized DL methods show great potential compared to traditional data fusion approaches (Hong et al., 2021).

In DL, three data fusion strategies can be distinguished (Schmitt and Zhu, 2016; Ramachandram and Taylor, 2017; Zhu et al., 2017): *observation-level fusion*, *feature-level fusion*, and *decision-level fusion* (DLF). *Observation-level fusion* combines different modalities into a common feature vector before being fed into the DL model, potentially hindering the identification of higher-level synergies between modalities. *Feature-level fusion* extracts modality-specific representations from the input, integrates them using a suitable fusion algorithm, and passes the result to the decision level. Multimodal representation and fusion components are optimized end-to-end, potentially uncovering beneficial higher-level multimodal relationships. *DLF* aggregates decisions from independent modality-specific models and is often preferred for its simplicity (Schmitt and Zhu, 2016; Ramachandram and Taylor, 2017; Baltrusaitis et al., 2019). *Heterogeneous data fusion* problems, where modalities differ substantially, typically involve fusion based on already abstracted information, such as extracted features or modality-specific model decisions (Hong et al., 2021).

Multimodal learning often faces scenarios of partially unavailable modalities, necessitating solutions to the *missing modality problem* (Mena et al., 2024; Kieu et al., 2024). This study addresses the prevalent scenario in which remote sensing data provide spatial continuity and comprehensive coverage, whereas SLI data remain incomplete due to buildings being unrecorded or obscured (Srivastava et al., 2019; Aravena Pelizari et al., 2021; Biljecki and Ito, 2021). Furthermore, obtaining complete and up-to-date SLI coverage across large areas is prohibitively expensive, particularly given the spatio-temporal dynamics of urban environments.

1.2. Related works

With regard to the geo-image modalities used in this study, Srivastava et al. (2019) and Hoffmann et al. (2019) employ VHR optical remote sensing data and SLI through feature-level fusion within a multi-stream CNN to predict urban land use classes of OpenStreetMap buildings. Srivastava et al. (2019) concatenate modality-specific feature vectors immediately before classification, while Hoffmann et al. (2019) additionally assess earlier-stage concatenation and DLF. Overall, both studies report significant accuracy gains employing both modalities. In Hoffmann et al. (2019) decision-level fusion in the majority of cases outperforms feature-level fusion. Srivastava et al. (2019) also address the inference of data points with missing SLI. Specifically, they employ the features of the available residual modality to project these data points, along with those containing SLI, into a common embedding

space and use the SLI features of their nearest neighbors in this space as substitutes.

Several recent studies propose methods for synergistic multimodal image classification of remote sensing and ground-based imagery based on the benchmark datasets AiRound (11 land use classes) and CV-BrCT (9 land use classes) introduced by Machado et al. (2021). Machado et al. (2021) compare multi-stream CNN models with early feature-level fusion and DLF variants. DLF through the multiplication of modality-specific class probability outputs is found to yield the highest accuracies. Machado et al. (2023) propose a retrieval CNN to replace missing modalities with similar existing samples from the database. Zhao et al. (2024) propose a teacher-student model to extract cross-modal knowledge and address partially missing data. Furthermore, they implement a cross-view attention module to capture correlations among the multimodal representations, outperforming both feature concatenation and DLF.

Chen et al. (2022) classify urban villages using SLI and VHR optical remote sensing data, also employing an attention based fusion to adaptively weight ground-based and top-view representations. Hosseinpour et al. (2022) beneficially leverage adaptive gating in the DL-based fusion of digital elevation models and VHR optical data for building segmentation.

The presented studies emphasize two key challenges of multimodal DL, *i.e.*, developing tailored fusion strategies to exploit positive synergies and the handling of missing modalities.

1.3. Conceptualization and contributions

Against the provided background, this research addresses the synergistic fusion of heterogeneous multimodal geospatial image data (SLI, VHR optical data and an nDSM with 2 m geometric resolution) for the vulnerability-related multicriteria characterization of buildings exposed to natural hazards (Fig. 1).

The inference of building attributes is addressed via DL-based image classification (Rawat and Wang, 2017), commonly referred to as scene classification in remote sensing (Cheng et al., 2020). This enables the comprehensive integration of spatial context information, which has proven effective for assigning complex semantic classes in urban environments using VHR remote sensing data (Herold et al., 2003; Geiß et al., 2015; Huang et al., 2018; Zhang et al., 2018; Martins et al., 2020). With respect to the addressed spatial entities, the proposed approach aligns with Huang et al. (2018), Zhang et al. (2018), Martins et al. (2020) and Wang et al. (2021), where representative sample locations are defined within previously delineated target objects. Restricting the application of the DL model to image patches extracted at such specifically defined data points, considerably reduces training data annotation efforts as well as the required number of model updates and predictions, increasing overall efficiency compared to pixel-based methods (Martins et al., 2020; Wang et al., 2021).

The base entities constitute building object polygons that can be derived from the input remote sensing data itself, *e.g.*, through instance segmentation (Stiller et al., 2019) or semantic segmentation (Neupane et al., 2021). Furthermore, OpenStreetMap provides crowdsourced building polygons, while Microsoft and Google published extensive building footprint data extracted from VHR remote sensing data[1] [2]. Focusing on delineated building objects, residual areas are excluded from the outset. This aligns with the SLI data (Section 2.1.1), which capture building façades but omit other urban elements.

Unlike remote sensing data with inherent geographic alignment, SLI exhibits spatial discrepancies between image content and recorded coordinates, as the latter represent camera positions rather than captured views (Qiao and Yuan, 2021). Inspired by Huang et al. (2018), Zhang et al. (2018), and Martins et al. (2020), we propose a method utilizing morphological line representations for representative spatial assignment and sampling within building objects. This integration process yields consistently localized SLI and remote sensing image patches, which serve as inputs for classification.

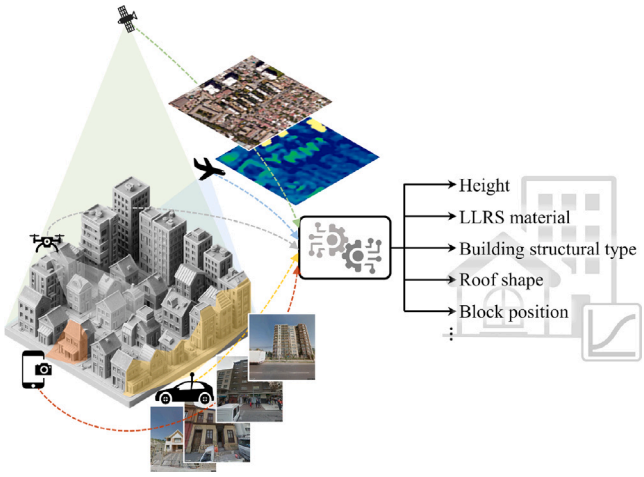

Fig. 1. Heterogeneous multimodal geospatial image data and deep multimodal multitask learning for the vulnerability-related multi-criteria characterization of buildings exposed to natural hazards.

To infer multiple vulnerability-relevant target variables from multimodal imagery, we propose a *multimodal multitask classification* (M^3TC) framework. It employs a feature-level fusion module—termed *task-wise modality attention* (TMA)—to optimally exploit synergies among input modalities by weighting their representations according to the specific requirements of each target task. In contrast to prior studies (Section 1.2), our approach provides a robust solution for multicriteria building characterization. From an application perspective, the (M^3TC) framework is designed to efficiently support the generic inventorization of exposed buildings, as envisioned in the faceted GED4ALL multi-hazard building taxonomy (Silva et al., 2022).

Many multimodal learning approaches enable synergistic inference despite missing modalities by leveraging dependencies within complete multimodal image data available during training (Baltrusaitis et al., 2019; Kieu et al., 2024). Examples include cross-modal image retrieval (Srivastava et al., 2019; Machado et al., 2023), cross-modality learning (Hong et al., 2021), and information exchange via cross-modal loss functions (Xie et al., 2023). However, the explicit consideration of spatio-contextual dependencies generally remains unconsidered. In contrast, this work addresses the challenge of missing SLI data using a transformer-based *SLI spatial context encoder*, which adaptively learns spatio-contextual representations from the façade views of the K nearest neighbors with available SLI data as substitutes. Consistent with Tobler's First Law of Geography (Tobler, 1970), which states that spatial interdependencies are stronger among proximate objects than distant ones, this approach leverages spatial correlations in the structural and visual properties of buildings. These correlations aim to capture distinctive patterns shaped by urban growth history, past natural disasters, evolving construction designs and regulations, as well as socio-economic factors such as demographics, income levels, and urban planning.

In summary, the technical innovations of this study are threefold: *(i)* the spatial integration of multimodal geo-image data via morphological line representations of building objects; *(ii)* TMA for optimizing data fusion in multimodal multitask learning; and *(iii)* the *SLI spatial context encoder* to mitigate missing SLI—together enabling the efficient area-wide extraction of reliable, faceted building exposure information.

[1] Microsoft, GlobalMLBuildingFootprints: https://github.com/microsoft/GlobalMLBuildingFootprints.

[2] Google, Open Buildings: https://sites.research.google/open-buildings/.

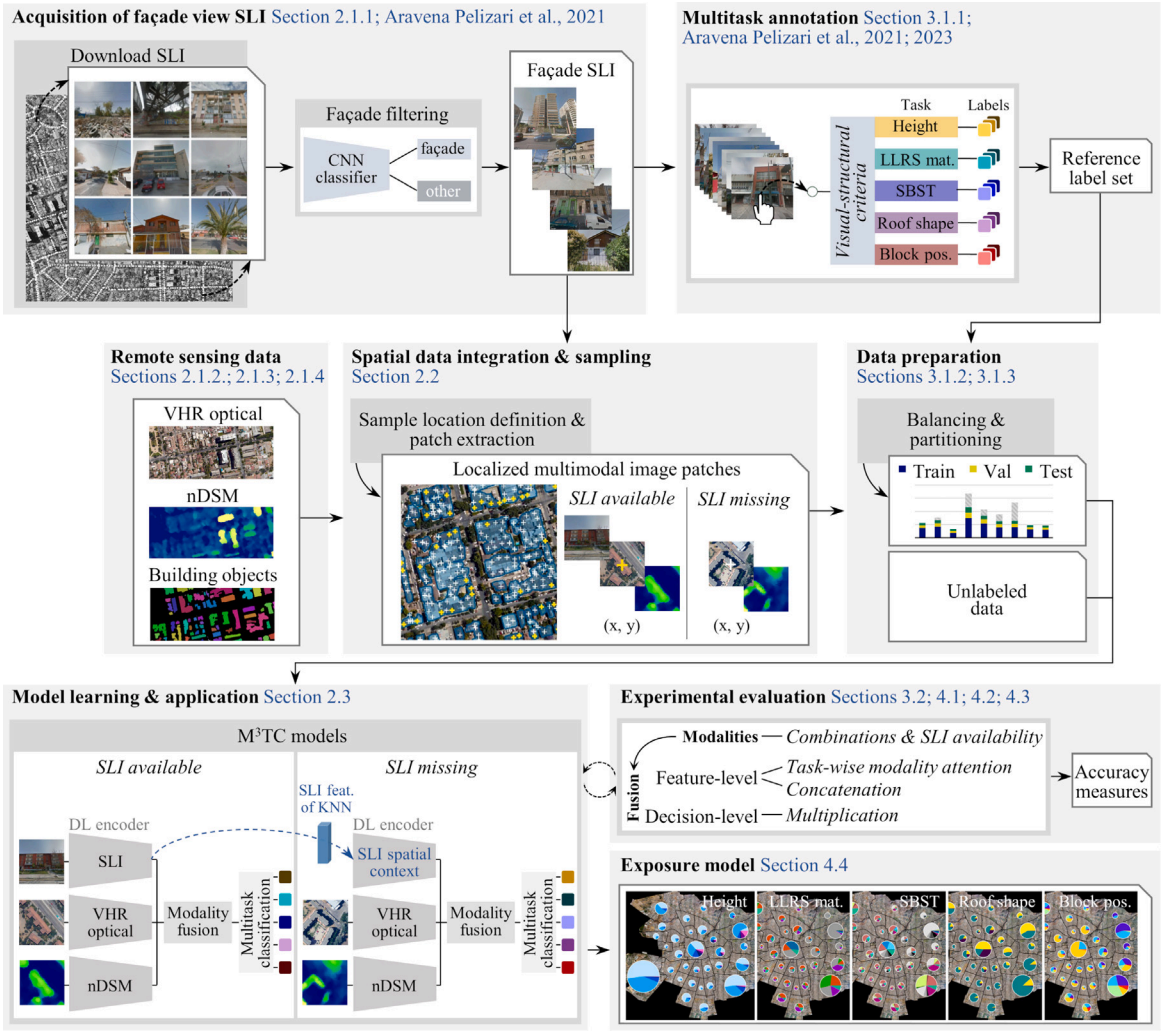

Fig. 2. Overview of the input data and processing steps. Details are provided in the indicated sections.

The presented M³TC framework is applied and experimentally evaluated for multi-criteria building characterization, focusing on five vulnerability-relevant attributes: *height*, *LLRS material*, *SBST*, *roof shape*, and *block position* (Section 3.1.1). The test site is Santiago de Chile, a city highly prone to earthquakes. Considering the two data scenarios—*SLI available* and *SLI missing*—the conducted experiments involve an extensive cross-comparison of generalization capabilities across individual geo-image modalities and their combinations. This includes a detailed assessment of the contributions from the TMA data fusion strategy and the *SLI spatial context encoder*. Finally, the most accurate models are used to generate a spatially continuous exposure model.

The remainder of the paper is organized as follows: Section 2 details the data utilized and the proposed methodology. Section 3 outlines the experimental setup, while Section 4 presents and discusses the results. Finally, Section 5 concludes the paper. Fig. 2 provides an overview of the input data and processing steps applied in this study, along with references to the corresponding paper sections.

2. Materials and methods

2.1. Data

2.1.1. Street-level imagery

The employed SLI data comprise GSV building façade views from Aravena Pelizari et al. (2021) within Santiago de Chile's 7M inhabitant metropolitan area: *(i)* scenes with a viewing direction perpendicular to the driving direction of the recording vehicle were sampled in a spatially stratified manner; *(ii)* a filtering procedure based on a *Places365* (Zhou et al., 2018) pretrained CNN separated façade from non-façade views. Example façade views are shown in Fig. 7.

2.1.2. VHR optical remote sensing data

With regard to VHR optical data, an RGB orthophoto mosaic from an airborne sensor with a geometric resolution of 0.4 m is utilized (Fig. 3a, b; IDE, 2015). Collected in January 2014, these data approximately align with the acquisition time of the SLI data. They are representative of VHR satellite imagery produced by modern multispectral systems such as WorldView-3 (0.31 m), GeoEye-1 (0.41 m), Pléiades-1 A and 1B (0.50 m), SkySat (0.50 m), and others.

2.1.3. Normalized digital surface model

In addition, a normalized digital surface model (nDSM) with a 2 m spatial resolution, derived from panchromatic tri-stereo imagery captured by the SPOT-7 satellite in 2014, is used (Fig. 3c). The original sensor data were processed by Stiller et al. (2021), which included the generation of a digital surface model (d'Angelo and Reinartz, 2011), from which the nDSM was derived (Perko et al., 2015). An accuracy assessment based on a very high resolution nDSM reported a mean absolute error of 2.9 m. A cost-effective alternative to classical photogrammetrically derived height information lies in DL-based height predictions from a single image (*monocular height estimation*; *e.g.*, Chen et al., 2023; Müller et al., 2023).

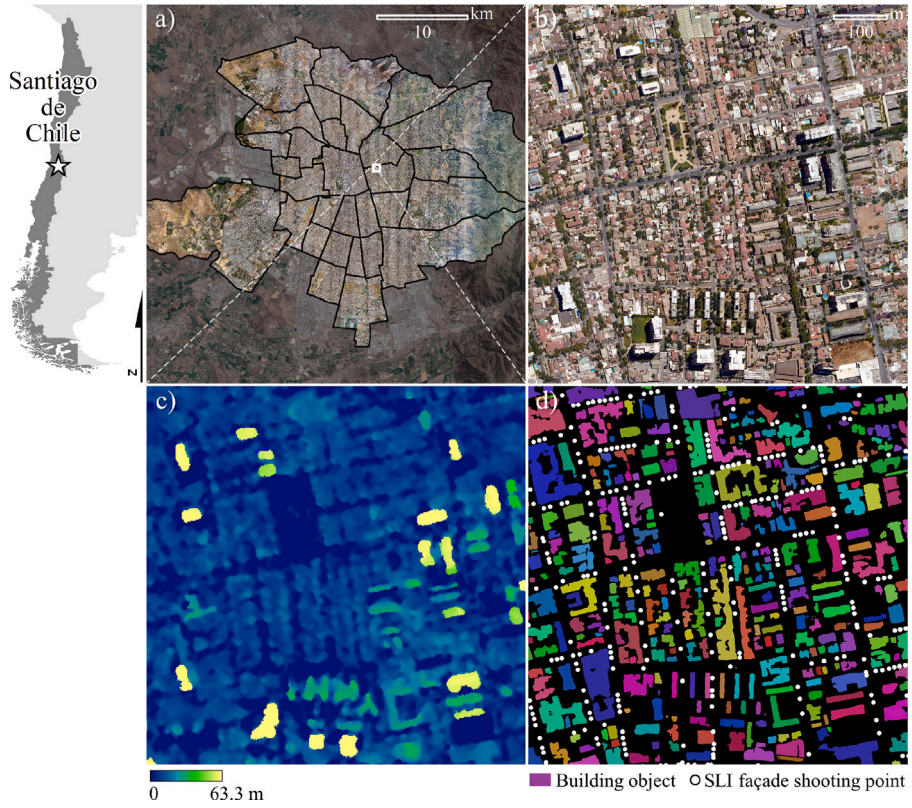

Fig. 3. Multimodal geospatial image data: (a) municipalities of Santiago de Chile (*Comunas*) with data coverage (shaded backdrop: Sentinel-2 image); (b) VHR optical remote sensing data; (c) nDSM; (d) building objects and street-level façade shooting positions.

2.1.4. Building object instances

Building objects extracted from the VHR optical remote sensing data as part of Stiller et al. (2019) using Mask RCNN instance segmentation (He et al., 2017) serve as geographic base entities for the exposure model (Fig. 3d). The dataset includes detailed delineations of built-up areas in the Santiago de Chile metropolitan area, with a spatial resolution ranging from individual buildings to building blocks. Freestanding buildings are captured individually, while building blocks represent dense, contiguous developments. The estimated accuracy of the building layer is 80.38% Intersection over Union, 92% Overall Accuracy (OA), and features a kappa (κ) of 0.83 (Stiller et al., 2019).

2.2. Spatial data integration

The spatial integration of the multimodal geospatial image data is based on the building object instances and comprises the following steps: *(i)* the extraction of morphological line representations of the building objects (hereafter referred to as *sample lines*), *(ii)* the localization of the street-level façade views on the sample lines, *(iii)* the definition of sample locations along the sample lines.

The notion behind the sample line is to delineate spatial locations within the building objects that correspond to their façade view at street level, either along their main axis or the object parts that face the street. The derivation of the sample line is shown in Algorithm 1 and in Fig. 4. The building objects are skeletonized to extract their main axes using the algorithm of Lee et al. (1994). Furthermore, the building objects are eroded by the distance d, set to 5 m, considering the building morphology of Santiago de Chile and corresponding object representations. Subsequently, the skeleton line sections overlapping with the eroded building object areas are erased (Fig. 4a). The final sample lines (Fig. 4b) are constructed from the remaining skeleton line sections and the outlines of the eroded building object areas. Consequently, building objects or parts of building objects with a width of ≤10 m are represented by their main object axis, while those with a width of >10 m are represented by their 5 m inwardly offset boundary.

Algorithm 1: Sample line (SL) extraction

procedure GET_$SL(objekt_i, d)$ ▷ Input: $Object_i$, distance d
  $S_i \leftarrow$ Skeletonize$(objekt_i)$ ▷ Skeletonize object
  $E_i \leftarrow$ Erode$(objekt_i, d)$ ▷ Erode object by distance d
  $SEr_i \leftarrow$ Erase(S_i, E_i) ▷ Erase skeleton by eroded object
  $SL_i \leftarrow$ Union$(SEr_i, Outline(E_i))$ ▷ Union of remaining object skeleton and eroded object outline
  return SL_i ▷ Output: Sample line
end procedure

The localization of available façade views within the building objects is performed as follows: a *link line* is generated from the coordinates of the façade shooting points and the horizontal viewing direction of the camera sensor, which underlies the façade view (Fig. 4b). The assignment location (M_{ALL} location) of a façade view is then defined as the point on the sample line closest to the first intersection between the link line and the boundary of the corresponding building object. At these M_{ALL} locations, all geo-image modalities are available for building characterization. Fig. 4e provides an overview of the M_{ALL} location coverage in the center of Santiago de Chile.

For the characterization of building objects or object parts beyond the data points with SLI coverage, sample locations are defined along the sample line at 12 m intervals. Building objects with a sample line shorter than 12 m are represented by their center point. These data points (M_{RS} locations; Fig. 4c, d) are captured exclusively by the remote sensing data.

Sub-object-level spatial assignment and sampling address densely built-up areas where individual buildings are challenging to delineate, instead forming parts of building block objects (Kraff et al., 2020).

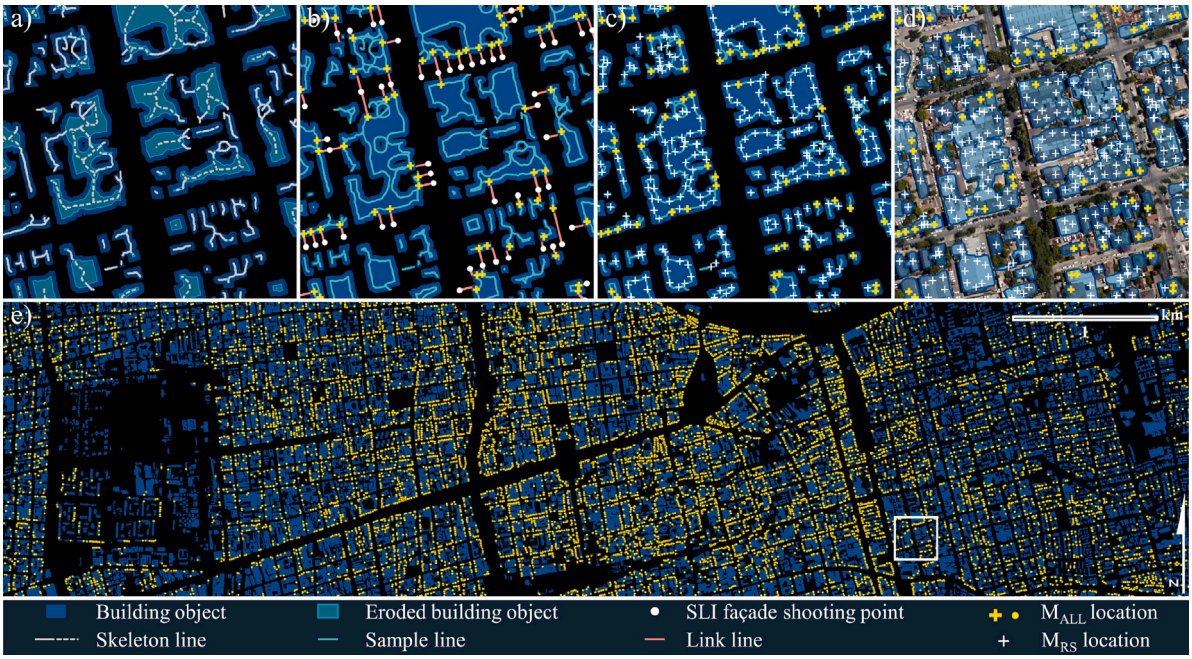

Fig. 4. Spatial integration of SLI and remote sensing data. (a) Sample line (SL) extraction: 1. Object *skeletonization*, 2. Object *erosion*, 3. *Erase-Union* operation (Algorithm 1); (b) The resulting SL and localization of street-level façade views using the link line, where coverage with all image modalities (M_{ALL} locations) is present; (c) Generation of sample locations in SL sections without SLI coverage, *i.e.*, where only remote sensing image modalities are available (M_{RS} locations); (d) Final sample locations; (e) West-East transect in the center of Santiago de Chile illustrating the density of M_{ALL} locations.

These objects may consist of buildings with varying vulnerability characteristics. This approach facilitates detailed vulnerability mapping even in high-density urban environments.

Using the identified sample locations as center points, corresponding remote sensing data patches are generated with dimensions of 88 × 88 m (*i.e.*, 220 × 220 pixels for the optical data and 44 × 44 pixels for the nDSM). This patch size enables the representation of buildings with varying footprint extents while capturing both the immediate spatial context and broader urban morphologies in the downstream feature encoding.

2.3. Deep multimodal multitask learning

The *multimodal multitask classification* (M^3TC) problem can be defined as follows: Given instances represented by M heterogeneous modalities $X^{MM} = \{X_m\}_{m=1}^{M}$, where $X_m \in \mathbb{R}^{d_m}$ denotes a d-dimensional instance representation space, each instance $x^{MM} \in X^{MM}$ is associated with a label space $Y^{MTC} = \{Y_t\}_{t=1}^{T}$, where $Y_t \in \{y_{t,c}\}_{c=1}^{C_t}$. Here, T denotes the total number of addressed classification tasks, and C_t corresponds to the task-specific number of classes. The goal of M^3TC ist to learn a prediction model $M^{M^3TC}(x^{MM}) : X^{MM} \rightarrow Y^{MTC}$, minimizing a joint loss over all tasks (Section 2.3.4).

The presented M^3TC architectural framework (Fig. 5) consists of an encoder module for each modality (*Multitask feature extraction*; Sections 2.3.1, 2.3.3), followed by a feature fusion and a classification module. The encoders utilize *hard parameter sharing* multitask learning to simultaneously learn shared initial feature representations, jointly optimized for inferring the target variables (*e.g.*, Aravena Pelizari et al., 2023). The proposed feature fusion module adaptively weights these representations according to the requirements of the individual target tasks (*task-wise modality attention fusion*; Section 2.3.2). For classification resulting multimodal fusion features are passed to a fully connected layer with *softmax* activation each. Final class labels are obtained from the softmax outputs s via $y_t = \arg\max_{t,c}(s_{t,c})$.

2.3.1. Image data encoders

The CNN based geo-image encoders were chosen for their parameter efficiency. For feature extraction from the SLI, *EfficientNetV2-B2* (Tan and Le, 2021) designed for input image data of size 260 × 260 pixels is employed. The EfficientNet architecture has already demonstrated high predictive accuracy in the context of multi-criteria building characterization based on SLI (Aravena Pelizari et al., 2023). For the remote sensing image modalities, CNN based on the DenseNet architecture (Huang et al., 2017) are utilized. Compared to other state-of-the-art methods, DenseNets have recently shown to be highly competitive in terms of accuracy and parameter efficiency across various VHR remote sensing multi-class scene classification benchmark datasets (Dimitrovski et al., 2023). The core elements of DenseNets are the *dense blocks*, which feature direct connections from each layer to all subsequent layers by concatenating their outputs. This optimizes feature reuse and information flow (Huang et al., 2017). A DenseNet with 120 convolutional layers (*DenseNet120*) is used for feature extraction from optical remote sensing data, and a DenseNet with 38 convolutional layers (*DenseNet38*) is used for feature extraction from nDSM data (Table 1). Shared multitask feature extraction from the input image data is concluded with *global average pooling* to keep subsequent data fusion and classification sparse.

2.3.2. Task-wise modality attention fusion

From the M^3TC setting, the following two hypotheses emerge: *(i)* not every available modality is equally relevant for deriving the various target variables being addressed, and *(ii)* not every feature representation from the shared encoders provides equal value for each task. To counteract potential accuracy losses due to these issues, the *task-wise modality attention* (TMA) fusion module is employed. Inspired by channel-wise feature weighting in *Squeeze-and-Excitation* blocks (Hu et al., 2018) and the use of underlying mechanisms in multimodal data fusion (Hosseinpour et al., 2022; Chen et al., 2022), TMA fusion involves task-specific, attention-gate-based weighting of multimodal input representations, followed by feature reduction. For a given set of

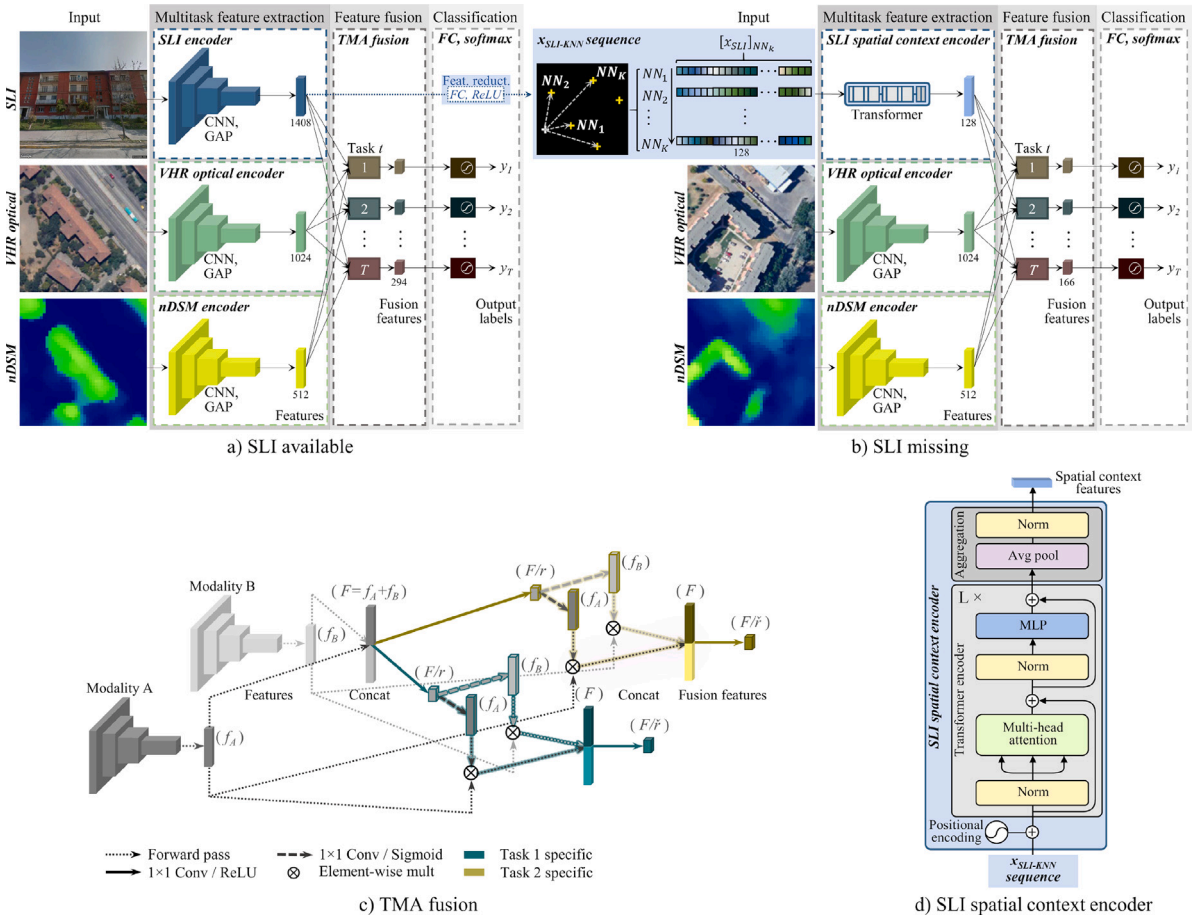

Fig. 5. *M^3TC model architectures*. Modality-specific encoders, shared across T classification tasks, enable multitask feature extraction. Imagery is processed via CNN followed by global average pooling (GAP; Section 2.3.1). *Task-wise modality attention* (TMA) fuses the features by weighting representations according to task requirements (Section 2.3.2). Fully connected (FC) layers with *softmax* activation perform task-wise classification (Section 2.3). (a) Configuration with SLI available. (b) Configuration with SLI missing: reduced SLI features from the K nearest neighbors (KNN) with available SLI are ordered by distance ($x_{SLI\text{-}KNN}$ *sequence*), and passed to the transformer-based *SLI spatial context encoder* (Section 2.3.3). (c) TMA fusion for two input modalities and two classification tasks, with relative feature vector sizes in brackets. (d) The *SLI spatial context encoder* in detail.

Table 1
DenseNet-encoder-architectures. *conv* indicates convolution-batch normalization (BN)-ReLU in the first layer; BN-ReLU-convolution thereafter.

	VHR optical encoder (input size: 224 × 224)		nDSM encoder (input size: 44 × 44)	
Layers	DenseNet120	Output size	DenseNet38	Output size
Convolution	7 × 7 conv, stride 2	112 × 112	7 × 7 conv, stride 2	22 × 22
Pooling	3 × 3 max pool, stride 2	56 × 56	3 × 3 max pool, stride 2	11 × 11
Dense Block (1)	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times6$	56 × 56	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times6$	11 × 11
Transition Layer (1)	1 × 1 conv 2 × 2 avg pool, stride 2	56 × 56 28 × 28	1 × 1 conv 2 × 2 avg pool, stride 2	11 × 11 5 × 5
Dense Block (2)	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times12$	28 × 28	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times12$	5 × 5
Transition Layer (2)	1 × 1 conv 2 × 2 avg pool, stride 2	28 × 28 14 × 14	–	–
Dense Block (3)	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times24$	14 × 14	–	–
Transition Layer (3)	1 × 1 conv 2 × 2 avg pool, stride 2	14 × 14 7 × 7	–	–
Dense Block (4)	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times16$	7 × 7	–	–
Feature aggregation	7 × 7 global avg pool	1 × 1	5 × 5 global avg pool	1 × 1

tasks T and feature representations from M available modalities, the TMA fusion process can be formulated as:

$$\begin{aligned} z &= \text{Concat}\left(x_1, \ldots, x_M\right), \quad z \in \mathbb{R}^{1\times1\times F_z} \\ \check{z}_t &= \text{ReLU}\left(\text{Conv}\left(z\right)\right), \quad t \in \{1, \ldots, T\}, \quad \check{z}_t \in \mathbb{R}^{1\times1\times\frac{F_z}{r}} \\ w_{x_{m,t}} &= \sigma\left(\text{Conv}\left(\check{z}_t\right)\right), \quad w_{x_{m,t}} \in \mathbb{R}^{1\times1\times F_{x_m}} \\ \check{x}_{m,t} &= w_{x_{m,t}} \otimes x_m \\ \check{x}_t &= \text{ReLU}\left(\text{Conv}\left(\text{Concat}\left(\check{x}_{1,t}, \ldots, \check{x}_{M,t}\right)\right)\right), \quad \check{x}_t \in \mathbb{R}^{1\times1\times\frac{F_z}{\check{r}}}, \end{aligned} \tag{1}$$

where *Concat* stands for vector concatenation, *Conv* for 1 × 1 convolution, F_j for the number of feature channels of vector j, σ for sigmoid activation and $\otimes$ for element-wise multiplication. The terms r and $\check{r}$ denote dimensionality reduction factors. $w_{x_{m,t}}$ represents the adaptively learned task-specific weight vectors for each modality, $\check{x}_{m,t}$ the weighted feature vectors, and $\check{x}_t$ the fused multimodal representations optimized to meet the requirements of the respective target tasks. A schematic visualization of TMA fusion for two modalities and two target tasks is shown in Fig. 5c.

2.3.3. SLI spatial context encoding

To address the issue of missing SLI data—such as at the M_{RS} locations (Fig. 4d)—the *SLI spatial context encoder* is proposed (Fig. 5b, d). This model learns spatial context representations from data points with available SLI to substitute the missing information. Specifically, the CNN-encoded SLI feature representations of the K nearest neighbors with SLI coverage, $\left[x_{SLI}\right]_{NN_k}$, $k \in 1, 2, \ldots, K$, are used to adaptively capture spatial interdependencies via a transformer. K is treated as a hyperparameter and optimized (Section 3.3). To mitigate the increase in input data size and model complexity as the number of nearest neighbors grows, the input SLI feature dimensionality is reduced via a fully connected layer with ReLU activation (Fig. 5a: *Feature reduction*). The resulting feature vectors are sorted in ascending order by distance to serve as input for spatial context encoding (Fig. 5b: $x_{SLI\text{-}KNN}$ *sequence*):

$$x_{\text{SLI-KNN}} = \begin{bmatrix} x_{\text{SLI}_1 NN_1} & \cdots & x_{\text{SLI}_{D_m} NN_1} \\ \vdots & \ddots & \vdots \\ x_{\text{SLI}_1 NN_K} & \cdots & x_{\text{SLI}_{D_m} NN_K} \end{bmatrix}. \tag{2}$$

The transformer relies on self-attention to capture intricate dependencies within the data (Vaswani et al., 2017), *i.e.*, *queries* (Q), *keys* (K), and *values* (V) represent different projections of the same input. It consists of L sequential blocks (see Fig. 5d), each comprising two modules: a *multi-head attention* (MHA) module and a multi-layer perceptron (MLP) module. The input and output of the modules are linked through residual connections (He et al., 2016). Unlike the configuration presented by Vaswani et al. (2017), this implementation applies layer normalization (Ba et al., 2016) to the inputs of the modules within the residual blocks promoting more stable training and faster model convergence (Xiong et al., 2020). As such, the output x^l of block l is computed as follows:

$$\begin{aligned} Q, K, V &= \text{Norm}\left(x^{l-1}\right) \\ x^{l\prime\prime} &= \text{MHA}\left(Q, K, V\right) + x^{l-1} \\ x^{l\prime} &= \text{Norm}\left(x^{l\prime\prime}\right) \\ x^{l} &= \text{MLP}\left(x^{l\prime}\right) + x^{l\prime\prime}, \quad x \in \mathbb{R}^{D_m\times N}. \end{aligned} \tag{3}$$

The MHA module consists of a predefined number of H scaled dot-product attention (SDA) layers (*heads*):

$$\text{SDA}_i\left(Q_i, K_i, V_i\right) = A_i V_i = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{D_t}}\right) V_i, i \in \{1, \ldots, H\}, \tag{4}$$

where A_i represents the attention weight matrix and $Q_i, K_i, V_i \in \mathbb{R}^{D_t\times N}$ matrices are independent trainable linear transformations of the input. The dot products of Q_i and K_i are scaled by $\sqrt{D_t}$ to mitigate vanishing gradients. MHA is derived concatenating all SDA layer outputs, followed by a projection back to the dimension of the original input D_m based on the weight matrix $W \in \mathbb{R}^{H\times D_t\times D_m}$:

$$\text{MHA}\left(Q, K, V\right) = \text{Concat}\left(SDA_1, \ldots, SDA_H\right) W. \tag{5}$$

The MLP module consists of two fully connected layers with a ReLU activation in between:

$$\text{MLP}\left(x^{l\prime}\right) = \text{ReLU}\left(x^{l\prime} W^1 + b^1\right) W^2 + b^2, \tag{6}$$

$W^1 \in \mathbb{R}^{D_m\times D_f}$, $W^2 \in \mathbb{R}^{D_f\times D_m}$, $b^1 \in \mathbb{R}^{D_f}$ and $b^2 \in \mathbb{R}^{D_m}$ are the associated trainable weight matrices and biases.

Before being passed to the transformer, sinusoidal position encodings (PE) are added to the elements of the input matrices to incorporate positional information (Vaswani et al., 2017). The $K\times N$ output feature matrix of the transformer is aggregated via *Average Pooling* (AP) across the K-axis and normalized after Ba et al. (2016). Correspondingly, with $\mathcal{T}$ denoting the transformer (Eqs. (3)–(6)), SLI spatial context representations (x_{CTX}) are obtained by:

$$x_{\text{CTX}} = \text{Norm}\left(\text{AP}\left(\mathcal{T}\left(\text{PE} + x_{\text{SLI-KNN}}\right)\right)\right). \tag{7}$$

Provided that *(i)* the spatial distribution of available data captures relevant spatio-contextual interdependencies, and *(ii)* data points exhibit distinctive features supporting their inference, the proposed approach can also mitigate the limited availability of spatial data modalities beyond SLI.

2.3.4. Optimization

During training, given a labeled example $\left(\left\{x_m^i\right\}_{m=1}^M, \left\{y_t^i\right\}_{t=1}^T\right)$, an M^{3}TC model learns by updating the shared and the task-specific parameters to jointly minimize categorical cross entropy for each task. The mulitask loss (L_{MTC}) is defined as the sum of all task-specific losses (L_t):

$$L_{\text{MTC}} = \sum_{t=1}^{T} L_t. \tag{8}$$

3. Experimental setup

3.1. Data: target variables, balancing, partitioning and quantities

3.1.1. Target variables

Considering the SLI and VHR optical data, along with an ontology based on visually inferable criteria (*visual-structural criteria*) jointly developed by local structural engineers and experienced image analysts, 24,263 data points within the study area were labeled according to five vulnerability-relevant target variables (Aravena Pelizari et al., 2021, 2023), *i.e.*: *(i)* the material type of the LLRS (MatLLRS), *(ii)* building height (number of storeys), *(iii)* a seismic building structural type (SBST), characterizing the main-load bearing system from a seismic vulnerability perspective, *(iv)* roof shape (RoofShp) and *(v)* block position (BlockPos), referring to a building's or dwelling unit's location relative to its neighbors. Labels denote the central building entity in the façade view. If multiple buildings are present, the label refers to the fully depicted building with the largest area share. If no building is fully captured, the label refers to the one with the largest visible area. Table 2 provides an overview on all target variables and associated class labels. Schematic exemplifications are shown in Fig. 6. In addition, Fig. 7 provides annotated façade views to visualize label manifestations.

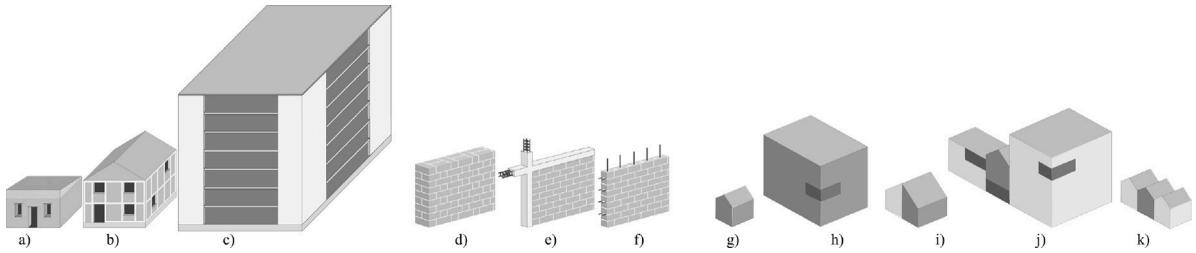

Fig. 6. Exemplification of addressed target variables (a-c: height, material LLRS, seismic building structural type, roof shape; d-f: details on masonry LLRSs; g-k: block position): (a) 1 storey, unreinforced masonry, MUR/H1, monopitch roof; (b) 2 storey, confined masonry, MCF/H1-2, pitched or gabled roof; (c) 5-7 storeys, reinforced concrete, CR/H5-7, flat roof; (d) unreinforced masonry wall; (e) confined masonry wall, *i.e.*, masonry with reinforced concrete confinement; (f) reinforced masonry wall, *i.e.*, masonry with steel bar reinforcement; (g) detached single-party; (h) detached multi-party; (i) semi-detached; (j) adjoining block development; (k) adjoining terraced.
Source: Modified after (Aravena Pelizari et al., 2023).

Fig. 7. Example façade imagery with class labels for the addressed target variables Height, MatLLRS, SBST, RoofShp, and BlockPos.

Table 2
Target building characteristics and multi-class manifestations (class numbers in brackets).

Height (6)	Material LLRS (7)	SBST (14)	Roof shape (3)	Block position (5)
1 storey (H1)	Unreinforced masonry (MUR)	MUR/H1	Flat (FLT)	Detached single-party (DET-SP)
2 storeys (H2)	Confined masonry (MCF)	MUR/H2-3	Pitched/ gabled (PIT-GAB)	Detached multi-party (DET-MP)
3-4 storeys (H3-4)	Reinforced masonry (MR)	MCF/H1-2	Monopitch (PIT-MON)	Semi-detached (SDET)
5-7 storeys (H5-7)	Reinforced concrete (CR)	MR/H1-2		Adjoining block development (ADJ-BD)
8-12 storeys (H8-12)	Wooden and non-engineered (W,UNK)	CR/H1-2		Adjoining terraced (ADJ-TR)
>13 storeys (H13+)	Other commercial and industrial (COM1,2,IND)	W,UNK/H1-2		
	Other office build. (COM3)	MCF/H3-4		
		MR/H3-4		
		CR/H3-4		
		CR/H5-7		
		CR/H8-12		
		CR/H13+		
		COM1,2,IND/H1+		
		COM3/H8+		

3.1.2. Balancing and partitioning

To mitigate class imbalance while accounting for the characteristics of multi-task annotated data—*i.e.*, interconnected task-specific class frequency histograms resulting from samples belonging to multiple classes—the reference data underwent label powerset-based random undersampling (LPRUS), as specified in Aravena Pelizari et al. (2023). The input data retention rate was set to 85%. Label powerset bins were also used when splitting the data into training, test, and validation sets (shares: 65%, 17.5%, 17.5%, respectively), ensuring representativity with respect to the occurring cross-task label combinations.

3.1.3. Quantities

The spatial data integration (Section 2.2) for the study area yields 161,474 M_{ALL} locations, where all geo-image modalities (*i.e*, SLI, opt, and nDSM) are available, as well as 1,281,460 M_{RS} locations, captured exclusively through remote sensing data (*i.e.*, opt and nDSM). As noted above, 24,263 of the M_{ALL} locations are labeled, representing the reference data for this study. The datasets resulting from data balancing and partitioning are shown in Fig. 8.

3.2. Experiments and validation

This study evaluates the potential of various geo-image modalities for vulnerability-related building characterization. Specifically, we investigate how classification accuracy can be improved by combining the available modalities through data fusion. Two cases are examined: *(i)* all modalities being available, as with M_{ALL} locations, and *(ii)* only VHR optical remote sensing data and an nDSM being available, but with missing SLI data, as with M_{RS} locations. Particular attention is given to assessing the contribution of spatial context modeling for accurate multi-criteria building characterization. Robustly handling both M_{ALL} and M_{RS} situations is crucial for deriving reliable, spatially continuous exposure models.

To assess the potential of the TMA method for multimodal image data fusion, the classic concatenation of encoded representations and decision-level fusion (DLF) serve as benchmarks. DLF is applied through

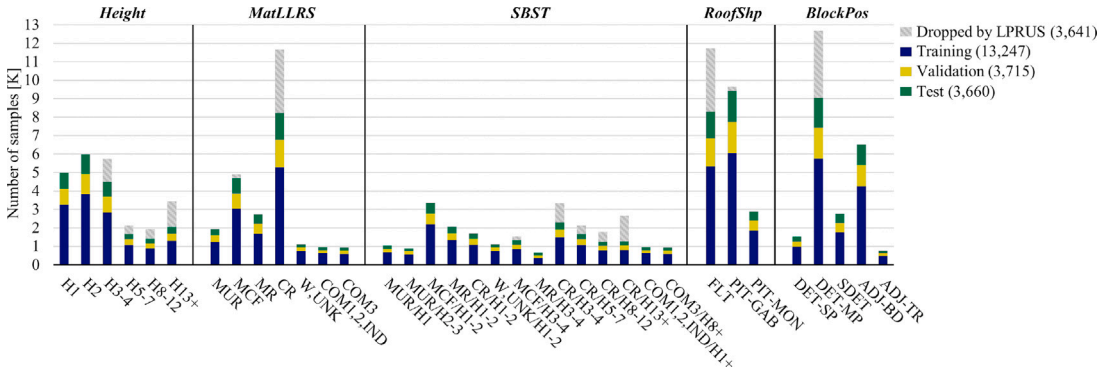

Fig. 8. Effect of LPRUS and class frequency distributions of training, validation, and test data for the five target tasks (total counts in parentheses).

the element-wise multiplication of softmax class probabilities resulting from M different modalities (*e.g.*, Machado et al., 2021) for the addressed classification tasks:

$$y_t = \arg\max_{t,c} \prod_{m=1}^{M} \left(s^m_{t,c} \right). \tag{9}$$

The transformer-based modeling of spatial context in the absence of SLI data is benchmarked using an additional variant of the *SLI spatial context encoder* that employs LSTM cells (Hochreiter and Schmidhuber, 1997) instead. Specifically, a stacked LSTM is implemented (*e.g.*, Rußwurm and Körner, 2020), comprising a bidirectional LSTM unit (Schuster and Paliwal, 1997) followed by a unidirectional unit to comprehensively capture spatio-contextual dependencies. The last hidden state is considered as spatio-contextual representation and passed on. For both variants of the *SLI spatial context encoder*, the influence of the number of nearest neighbors on classification accuracy is examined.

The generalization ability of the models is reported in terms of overall accuracy (OA), κ statistics, and F_1-scores, all derived from seven independent realizations. κ quantifies the agreement between multiclass predictions and reference labels, accounting for the agreement expected by chance (Cohen, 1960). Additionally, task-specific accuracy values for different modality combinations and the applied data fusion method (mf) are aggregated as cumulative residuals in accuracy relative to a defined reference modality (b):

$$\Delta^{mf}_b = \sum_{t=1}^{T} \left(\mathcal{M}_{mf,t} - \mathcal{M}_{b,t} \right) / \mathcal{M}_{b,t}. \tag{10}$$

$\mathcal{M}_{\cdot,t}$ referring to the measure used to assess the accuracy of task t.

3.3. Model parametrization and training

The reduction factors for TMA fusion, r and $\check{r}$ (Eq. (1)), are set to 16 and 10, respectively. Both values achieve a good balance between model accuracy and complexity, the former aligning with the findings of Hu et al. (2018).

The SLI representations for modeling spatio-contextual dependencies are based on the most accurate SLI encoder model from seven realizations. The dimensionality of the feature sets fed into the *SLI spatial context encoder* is defined as $D_m = 128$, and the number of transformer encoder blocks in the *SLI spatial context encoder* is set to $L = 3$. The intermediate dimension of the MLP modules is defined as $D_f = D_m \times 4$ (Eq. (6)). To prevent overfitting, dropout (Srivastava et al., 2014) is applied within the transformer blocks.

While the parametrization of the transformer-based *SLI spatial context encoder* remains unchanged across all experiments, tuning is performed for the LSTM-based variant on the dimensionality of the hidden states D_h. Specifically, the value pairs for the uni- and bidirectional LSTM units $\left(D_{h_{uni}}, D_{h_{bi}} \right) \in \{(64, 96), (128, 192), (192, 288)\}$ are considered. Dropout is applied between the two LSTM units.

For both the transformer- and LSTM-based *SLI spatial context encoders*, the number of nearest neighbors is varied within $K \in \{3, 5, 10, 25, 50, 75, 100, 150, 200, 300, 500\}$ to optimize the spatio-contextual representations. To ensure reliable estimates of generalization accuracy and avoid bias in the training process due to data overlap in the nearest-neighbor space, validation and test data must be excluded from the set of M_{ALL} locations considered as nearest neighbors.

For training, the CNN encoders are initialized with ImageNet (Russakovsky et al., 2015) pre-trained parametrization. Since we noticed that leveraging the full potential of ImageNet pre-training is beneficial for extracting information from the nDSM data, each patch is fed three times, simulating an RGB image. To stabilize the optimization process, the *SLI spatial context encoder* undergoes a separate warm-up phase, increasing the learning rate linearly from $1e{-}10$ to $1e{-}3$ over 413 update steps (*i.e.*, 1 epoch). The LSTM units are initialized using *Glorot uniform* initialization (Glorot and Bengio, 2010). The fully connected classification layers are *He normal* initialized (He et al., 2015) and subject to L2 weight regularization (L2 = $1e{-}4$). Thereon, all models are uniformly trained with *Adam* optimization (Kingma and Ba, 2014) and an initial learning rate of $1e{-}3$. For comprehensive yet efficient training, the learning rate is reduced by a factor of 0.1 when validation accuracy plateaus. Early stopping is applied to prevent overfitting. Considering the Nvidia RTX A4000 GPU's 16 GB memory, all models are trained with a batch size of 32.

4. Results and discussion

4.1. Performance: impact of input modalities and fusion strategy

Table 3 provides a comparative overview of the models' generalization capabilities based on the available geo-image modalities and the applied data fusion strategies under the two scenarios: *SLI available* (top section) and *SLI missing* (bottom section). It presents the mean estimated generalization accuracies (OA and κ), both task-specific and aggregated across all classification tasks. The absolute added value of incorporating additional modalities and the applied fusion method is indicated by the cumulative residuals of the task-specific accuracies relative to a reference modality (Δ^{mf}_b, Eq. (10)).

The results from individual modalities indicate that SLI data consistently delivers the highest classification performance across all tasks, with a substantial margin (mean $OA = 87.30\%$, mean $\kappa = 83.02\%$). This is followed by the accuracies obtained with VHR optical imagery (opt; mean $OA = 74.86\%$, mean $\kappa = 67.22\%$), which outperform the nDSM-based accuracies (mean $OA = 65.87\%$, mean $\kappa = 54.18\%$) for all tasks except height classification. In the absence of SLI, the highest mean task accuracies were achieved using spatio-contextual representations learned from the available SLI data (ctx; $OA = 77.42\%$, $\kappa = 70.37\%$). These results highlight both the high semantic information content of SLI and the potential of spatio-contextual information for the structural characterization of buildings.

Table 3
Mean accuracy values [%] for the target variables based on the availability of geo-image modalities and the fusion method, derived from seven independent runs. Top section: *SLI available* (M_{ALL} locations); bottom section: *SLI missing* (M_{RS} locations). In case of multimodal data fusion, cumulative residuals of task-specific accuracies relative to a reference modality $\left(\Delta_b^{mf}\right)$ are provided. Reference modality b is underlined and represents the modality with the highest individual accuracy within the combination.

	Data	Fusion	*Height*		*MatLLRS*		*SBST*		*RoofShp*		*BlockPos*		*Mean*		Δ_b^{mf} (*mean*)	
			OA	κ	OA	κ	OA	κ	OA	κ	OA	κ	OA	κ	OA	κ
SLI available	SLI	-	90.16	87.52	87.77	83.88	82.45	80.79	88.89	81.76	87.25	81.16	87.30	83.02	-	-
	SLI+opt	CNC	89.87	87.15	88.24	84.49	82.58	80.93	89.30	82.45	89.10	83.87	87.82	83.78	+3.0	+4.7
		DLF	89.86	87.13	88.29	84.57	83.01	81.41	89.31	82.46	89.27	84.12	87.95	83.94	+3.7	+5.6
		TMA	**90.38**	**87.81**	**88.43**	**84.73**	**83.01**	**81.40**	**89.70**	**83.14**	**89.73**	**84.82**	**88.25**	**84.38**	**+5.4**	**+8.3**
	SLI+nDSM	CNC	90.30	87.71	87.67	83.73	82.46	80.80	89.07	82.06	87.44	81.46	87.39	83.15	+0.5	+0.8
		DLF	90.45	87.89	88.09	84.28	82.96	81.35	89.10	82.12	87.49	81.49	87.62	83.43	+1.8	+2.4
		TMA	**90.80**	**88.33**	**88.38**	**84.66**	**83.43**	**81.86**	**89.31**	**82.45**	**87.80**	**81.97**	**87.94**	**83.85**	**+3.7**	**+5.0**
	SLI+opt+nDSM	CNC	90.15	87.51	88.26	84.49	82.88	81.26	89.58	82.89	89.33	84.23	88.04	84.08	+4.2	+6.5
		DLF	89.95	87.25	88.25	84.50	82.94	81.33	89.69	83.08	89.20	84.03	88.00	84.04	+4.0	+6.2
		TMA	**90.95**	**88.52**	**88.53**	**84.86**	**83.74**	**82.20**	**89.69**	**83.09**	**89.74**	**84.84**	**88.53**	**84.70**	**+7.1**	**+10.2**
SLI missing	ctx	-	74.88	68.27	76.79	69.31	67.04	63.87	82.87	71.88	85.50	78.53	77.42	70.37	-	-
	opt	-	70.03	61.93	74.57	66.12	60.98	57.22	82.98	71.94	85.74	78.90	74.86	67.22	-	-
	nDSM	-	70.13	62.08	61.41	47.85	51.30	46.24	74.12	56.10	72.46	58.62	65.88	54.18	-	-
	ctx+opt	CNC	**76.41**	**70.14**	78.70	71.78	69.17	66.27	**84.43**	**74.30**	87.46	81.42	79.23	72.78	+11.9	+17.1
		DLF	75.66	69.24	77.72	70.55	68.25	65.25	83.62	73.00	87.21	81.08	78.49	71.82	+7.0	+10.2
		TMA	76.27	69.99	**79.01**	**72.12**	**69.54**	**66.64**	84.33	74.13	**87.62**	**81.61**	**79.35**	**72.90**	**+12.7**	**+18.0**
	ctx+nDSM	CNC	77.89	72.08	77.85	70.66	69.73	66.85	83.24	72.40	86.85	80.47	79.11	72.49	+11.4	+15.4
		DLF	77.71	71.84	77.76	70.54	69.70	66.81	83.06	72.15	86.92	80.62	79.03	72.39	+10.9	+14.7
		TMA	**78.62**	**73.01**	**78.10**	**70.97**	**69.91**	**67.05**	**83.39**	**72.62**	**87.01**	**80.70**	**79.40**	**72.87**	**+13.4**	**+18.1**
	opt+nDSM	CNC	75.59	69.02	74.52	66.04	65.04	61.62	**83.14**	**72.09**	85.31	78.25	76.72	69.41	+14.2	+18.4
		DLF	75.09	68.35	73.71	65.03	64.41	60.94	82.91	71.75	85.84	79.04	76.39	69.02	+11.7	+15.1
		TMA	**77.28**	**71.14**	**75.14**	**66.86**	**66.71**	**63.41**	83.07	71.97	**86.29**	**79.66**	**77.70**	**70.61**	**+21.3**	**+27.8**
	ctx+opt+nDSM	CNC	78.52	72.85	78.42	71.38	70.21	67.38	83.93	73.50	87.61	81.61	79.74	73.34	+15.5	+21.4
		DLF	77.97	72.19	78.23	71.16	69.92	67.06	83.37	72.64	87.42	81.37	79.38	72.88	+13.2	+18.1
		TMA	**79.19**	**73.67**	**79.39**	**72.64**	**71.12**	**68.38**	**84.43**	**74.27**	**87.92**	**82.10**	**80.41**	**74.21**	**+19.9**	**+27.6**

In both data scenarios—where SLI is available and where it is absent—starting with a single modality, the integration of each additional modality leads to a substantial increase in accuracy. Among the evaluated fusion approaches, TMA fusion consistently achieves the highest accuracy compared to the benchmark methods feature concatenation (CNC) and DLF. Accordingly, the highest accuracies are achieved through the TMA fusion of all available modalities.

In the case of SLI availability, a mean task accuracy of up to 88.53% OA and 84.70% κ (SLI+opt+nDSM with TMA fusion) is achieved, corresponding to mean accumulated accuracy gains of up to +7.1% OA and +10.2% κ. It becomes apparent that the already high accuracies resulting from the street-level perspective can be considerably improved with the addition of top-view image modalities.

In the absence of SLI, substantially lower accuracy levels can be observed, with the highest mean task accuracy values reaching OA = 80.41% and κ = 74.21% (ctx+opt+nDSM with TMA fusion). Compared to the exclusive use of ctx representations, integrating the remote sensing-based modalities via TMA fusion results in mean task-accumulated gains of +19.9% in OA and +27.6% in κ. Envisaging the considered data fusion methods, this corresponds to a 6.2 percentage points (pp.) and 9.5 pp. higher gain in terms of κ accuracy compared to CNC and DLF, respectively.

When evaluating the accuracies of data fusion methods across different modality combinations, the added value of the TMA method becomes particularly evident in the opt+nDSM models. Relative to using opt alone, mean accumulated accuracy gains of +27.8% in κ are achieved—9.4 pp. higher than with CNC and 12.7 pp. higher than with DLF.

Hypothetically assuming the test dataset is representative of the entire study area, we extrapolated the misclassification difference between CNC and TMA fusion to all unlabeled data points (n = 1,418,671), based on the mean OA estimated for their respective data availability scenario (*i.e.*, SLI+opt+nDSM or ctx+opt+nDSM). This suggests that TMA fusion would reduce misclassifications by approximately 46,573 across all tasks.

Fig. 9 presents the results as boxplots of the cumulative accuracy residuals across tasks, relative to the model run representing the median of the mean task accuracies for a defined reference modality. Visualizing the stochastic variability in model training, this provides further insights into *(i)* the relative accuracy gains achieved through modality combinations and *(ii)* the impacts of different data fusion strategies. It becomes evident that the TMA fusion method consistently achieves better data fusion results in the majority of realizations compared to the benchmark methods CNC and DLF. Not included in Table 3, Fig. 9c shows the cumulative cross-task accuracy residuals in the absence of SLI when additional modalities are incorporated, starting from VHR optical data. The opt+nDSM-TMA combination yields a median increase in κ of +25.59%, opt+ctx-TMA +41.22%, and opt+nDSM+ctx-TMA reaches +51.28%. This highlights the potential of spatio-contextual information inferred from data points with available SLI to mitigate accuracy losses due to missing SLI. Moreover, it underscores the substantial accuracy gains in structural building characterization achieved through the fusion of the considered modalities.

Although TMA fusion models employ considerably more trainable parameters than CNC and DLF models—particularly when all modalities are included—the overall model sizes remain moderate. The SLI+opt+nDSM model with TMA fusion comprises 64.62M parameters (the largest model), while the ctx+opt+nDSM model with TMA fusion comprises 25.07M parameters (the 2nd largest model). Training and inference times across CNC, DLF, and TMA models were of similar magnitude. *E.g.*, mean per-epoch training times (*min:sec*) for the SLI+opt+nDSM models were 5:33 with CNC, 4:53 with DLF, and 4:45 with TMA; for the ctx+opt+nDSM models, they were 2:42 with CNC,

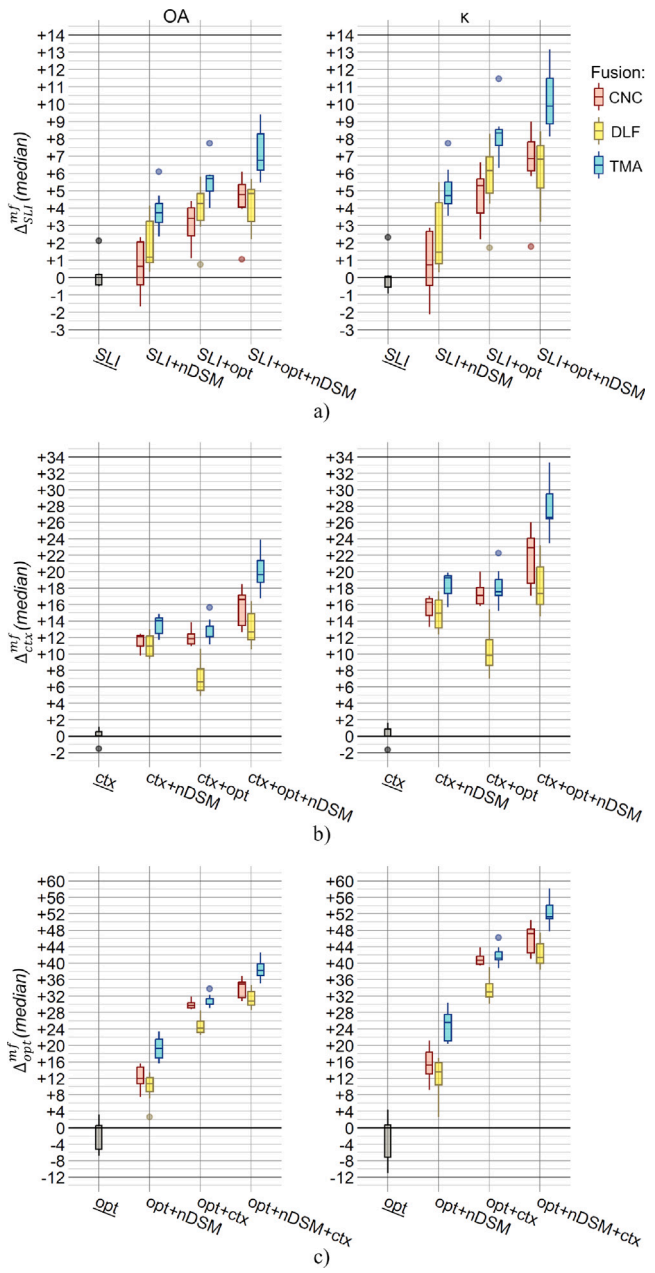

Fig. 9. Cumulative residuals of task-specific accuracies (OA and κ) for combinations of geo-image modalities and data fusion strategies, relative to the median values of different reference modalities [%]: (a) Δ^{mf}_{SLI}, (b) Δ^{mf}_{ctx}, (c) Δ^{mf}_{opt}.

2:47 with DLF, and 2:47 with TMA. The number of epochs required for convergence did not vary substantially. Inference times were likewise comparable (*min:sec* per 5k data points, batch size = 4): 1:07 with CNC, 1:09 with DLF, and 1:13 with TMA for the SLI+opt+nDSM models, and 0:50 with CNC, 0:50 with DLF, and 0:53 with TMA for the ctx+opt+nDSM models.

4.2. Insights on modeling spatio-contextual dependencies

Here, the modeling and integration of spatio-contextual representations (Section 2.3.3) for classifying data points with missing SLI coverage (*i.e.*, the M_{RS} locations; Section 2.2) are examined in greater detail.

First, the performance of classification models (task means of OA and κ) based solely on spatio-contextual information is rendered as a function of the number of considered nearest neighbors (Fig. 10a). It becomes evident that, even when considering only the SLI representations of the first nearest neighbors of the data points, a mean accuracy of 71.15% OA and 61.91% κ can be achieved. This is remarkable and already underscores the relevance of spatio-contextual information for the physical characterization of buildings. Using the *SLI spatial context encoder* with three nearest neighbors, classification accuracy can be increased by 4.52 pp. in OA and 6.02 pp. in κ, reaching 75.67% and 67.93%, respectively. As the number of considered nearest neighbors increases, accuracy continues to improve until convergence occurs at approximately KNN = 50 (OA = 77.29%, κ = 70.17%). As such, it is demonstrated that the proposed approach also effectively utilizes information from more distant neighbors to represent spatial context.

Analogously, Fig. 10b illustrates how accuracy evolves as the number of considered nearest neighbors successively increases when fusing spatio-contextual information from the SLI data with the remote sensing modalities opt and nDSM. Each model configuration employs the TMA data fusion method. The starting point is the exclusive use of remote sensing data (KNN = 0, *i.e.*, opt+nDSM), which results in a mean task accuracy of 77.70% OA and 70.61% κ. Next, the SLI features from the respective nearest neighbor are integrated as additional representations of the data points (KNN = 1, *i.e.*, opt+nDSM+NN-SLI feat.). As a result, classification accuracy slightly decreases, indicating that a single nearest neighbor does not yet provide sufficient complementary contextual knowledge to achieve a global benefit. Instead, it introduces disruptive noise into the model. For KNN ≥ 3, model performance results from the integration of learned spatio-contextual representations. The accuracies achieved with the proposed transformer-based *SLI spatial context encoder* (blue bars) are compared to those of an LSTM-based variant (yellow bars; Section 3.2). The integration of spatio-contextual representations considering multiple nearest neighbors leads to an improvement in model accuracy from the outset. Both the accuracies of opt+nDSM+ctx$^{\text{Transformer}}$ and opt+nDSM+ctx$^{\text{LSTM}}$ increase with higher KNN values, reaching a peak at KNN = 100 (OA = 80.41% and κ = 74.21% with transformer, OA = 80.05% and κ = 73.79% with LSTM). The mean accuracy of the opt+nDSM+ctx$^{\text{LSTM}}$ models is consistently outperformed by that of the opt+nDSM+ctx$^{\text{Transformer}}$ models. For KNN = 100, the mean cumulative task-specific accuracy residuals $\Delta^{\text{opt+nDSM+ctx}^{\text{Transformer}}}_{\text{opt+nDSM}}$ are 3.18 pp. higher in κ than $\Delta^{\text{opt+nDSM+ctx}^{\text{LSTM}}}_{\text{opt+nDSM}}$. As distances increase, the interdependencies between the visual appearances and the addressed structural characteristics of buildings diminish. Beyond a certain threshold, additional nearest neighbors no longer add value and instead introduce interference. Consequently, classification accuracy decreases once the accuracy peak is surpassed. However, this decrease is more pronounced for opt+nDSM+ctx$^{\text{LSTM}}$ models: While the OA and κ values for opt+nDSM+ctx$^{\text{Transformer}}$ models remain constant with KNN ≥ 200, these values continue to decrease for opt+nDSM+ctx$^{\text{LSTM}}$ models.

In summary: *(i)* The *SLI spatial context encoder* facilitates the modeling of meaningful spatio-contextual representations by leveraging SLI feature sequences from nearest neighbors within a spatial distance hierarchy; *(ii)* an adequate number of nearest neighbors is critical to fully harness the encoder's potential; and *(iii)* the proposed transformer-based variant outperforms its LSTM-based counterpart, delivering higher accuracy and greater robustness to long data sequences.

4.3. Interactions between input modalities and class-wise accuracies

Fig. 11 visualizes the class-specific accuracies (F_1-scores) of the target variables, resulting from the use of the different considered geo-image modalities and their combinations. For the latter, results are based on TMA data fusion.

Due to the already high SLI accuracies, the potential for class-wise improvements by additionally considering the remote sensing-based modalities is limited, but nevertheless clearly discernible (Fig. 11a). As expected, the inclusion of the nDSM in particular improves the

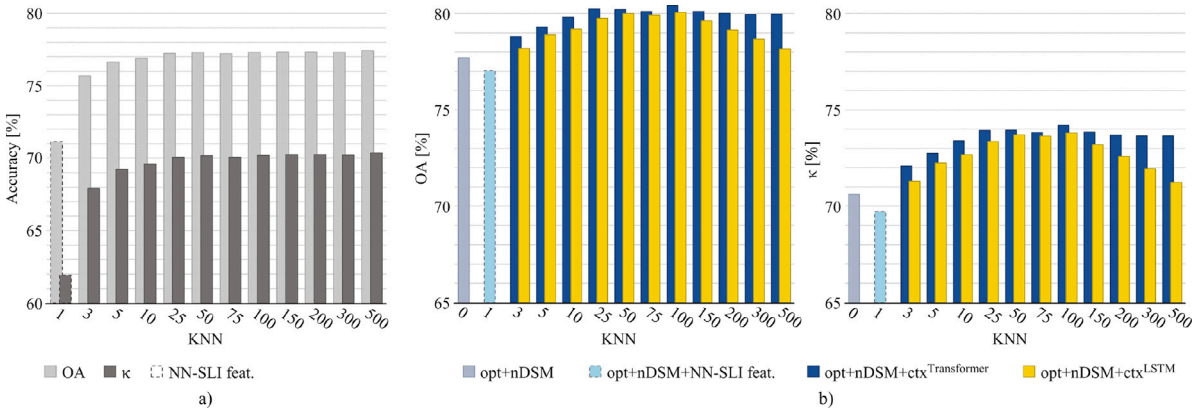

Fig. 10. Modeling and integration of spatio-contextual dependencies: (a) Accuracies of the transformer-based *SLI spatial context encoder* classifier as a function of the number of considered K nearest neighbors (KNN). In the case of KNN = 1 (dashed line), the SLI representations of the nearest neighbor are used for classification. (b) Accuracy evolution of the opt+nDSM+ctx-TMA model configuration and comparison of LSTM- and transformer-based context modeling (left OA, right κ). The starting point is the classification without considering spatial context (KNN = 0), *i.e.*, based solely on the remote sensing data (opt+nDSM). This is followed by the integration of the SLI representations of the nearest neighbor (KNN = 1) of a data point and the spatio-contextual representations modeled by the *SLI spatial context encoder* (KNN ≥ 3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

accuracy of classes correlated with building height. This effect is most pronounced in higher neighboring classes due to reduced omission and commission errors (*e.g.*, H5-7, H8-12, H13+ for the *Height* task and CR/H5-7, CR/H8-12, CR/H13+ for the *SBST* task). Additionally, other target variables, such as the *MatLLRS* classes MCF, MR, W, UNK, and COM3, the *RoofShp* class PIT-MON, as well as the *BlockPos* class ADJ-BD, benefit from nDSM integration. The fusion of SLI and top-view VHR optical data (opt) also meets expectations regarding improvements in prediction accuracy at the class level. Pronounced benefits are observed for the *MatLLRS* classes MR, COM1-2, and IND; the *SBST* classes MR/H1-2, CR/H1-2, and MCF/H3-4; the RoofShp class PIT-MON; as well as for the *BlockPos* task in general. *E.g.*, due to occlusion by fences, walls, or vegetation, SLI data alone cannot always unambiguously distinguish whether a building is a detached single-party house (DET-SP), a semi-detached house (SDET), or part of an adjoining block development (ADJ-BD). Additionally, the limited field of view of the SLI can hinder the determination of whether a multi-party building is part of an adjoining block development or a standalone structure (DET-MP). In such cases, top-view VHR optical images provide valuable complementary spatio-contextual cues. The integration of SLI data with both remote sensing modalities (SLI+opt+nDSM) generates additional synergies for most classes. Overall, this combination achieves the highest F_1 accuracy values for 19 of the 35 target classes (SLI+nDSM: 8/35; SLI+opt: 7/35; SLI: 1/35).

Fig. 11b presents the F_1-scores of data points where SLI is missing. Regarding the individual modalities, the nDSM data shows its highest potential for distinguishing height-related classes. Apart from the height- and *SBST*-classes with floor numbers H5-7, H8-12, and H13+, the accuracies achieved using VHR optical data substantially exceed those of the nDSM in most cases. The spatio-contextual representations (ctx) outperform both the VHR optical and nDSM data for most classes. The accuracies achieved through the combination of different geo-image modalities generally surpass those of individual modalities, even at the class level. Overall, ctx+opt+nDSM produces the most accurate model across all individual classes, yielding the highest F_1-scores for 25 out of the 35 target classes (ctx+opt: 4/35; ctx+nDSM: 3/35; ctx: 2/35; opt+nDSM: 1/35).

Following the global accuracy values (Table 3), the potential of individual geo-image modalities as well as the positive synergies due to their combination via TMA fusion are clearly evident at the individual class level, both when SLI data is available and when it is not. In the latter case, integrating the proposed ctx representations results in the highest F_1-scores for 34 of the 35 classes.

Additionally, Fig. 12 shows the normalized confusion matrices for the most accurate models identified in both situations: *SLI available* (Fig. 12a) and *SLI missing* (Fig. 12b), providing a class-specific overview of the nature of prediction errors.

4.4. The exposure model

To ultimately obtain the spatially distributed exposure models, considering the two data scenarios—*SLI available* and *SLI missing*—all sample locations are classified using the best-performing model, respectively (corresponding accuracy values are shown in Fig. 12). Starting from the base entities of the building objects, the derived point-based information layer can be aggregated into any larger geographical units (*e.g.*, broader building blocks, spatial grids, or administrative units).

The resulting exposure model for Santiago de Chile is presented in Fig. 13. From top to bottom, the building height, material of the lateral load resisting system (MatLLRS), seismic building structural type (SBST), roof shape (RoofShp), and block position (BlockPos) are shown. The left column displays the derived exposure information, aggregated to the administrative units of comunas. The center and right columns (zoom boxes A and B, respectively) show the aggregation to the building objects. The pie charts represent the class shares of the five target variables for the respective aggregation objects. The size of the diagrams corresponds to the number of data points considered. In total, the exposure model comprises more than 1.4 million classified data points. Spatial data assignment and sampling at the sub-object level (Section 2.2) enable the depiction of vulnerability-relevant characteristics as detailed distributions, even with respect to the building object level.

The spatially continuous exposure information reflects the heterogeneous spatial patterns characteristic of Santiago de Chile (Fig. 13, left column). The middle and right columns spotlight the high resolution of the exposure model and reveal that distinct spatial patterns and variability in the considered building characteristics prevail even at a small spatial scale. Accordingly, the buildings' vulnerability to natural hazards also exhibits distinct spatial variabilities, making such information highly valuable for spatial risk modeling (Aravena Pelizari et al., 2021; Gómez Zapata et al., 2021; Geiß et al., 2023).

The proposed data and methods enable the area-wide collection of vulnerability-relevant building attributes in an automated manner, offering a unique combination of spatial and thematic resolution. This is crucial for comprehensive multihazard risk analyses. The required spatial resolution of the exposure model depends on the extent, the

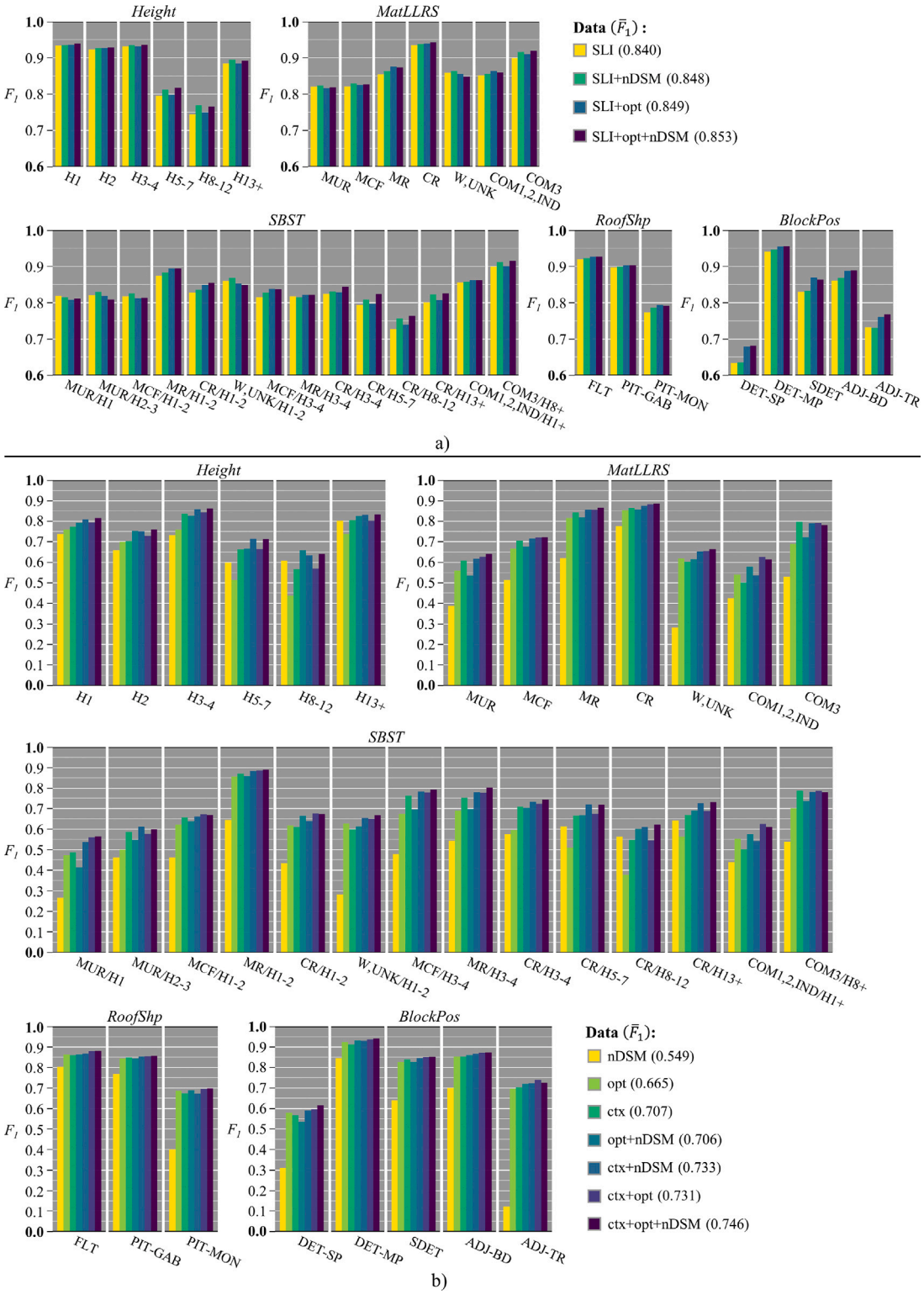

Fig. 11. Class-wise mean F_1-scores for the addressed classification tasks as a function of the input modalities: (a) *SLI available*; (b) *SLI missing*. When multiple input modalities are used, TMA data fusion is applied. Cross-task mean F_1-scores ($\bar{F}_1$) are provided in parentheses.

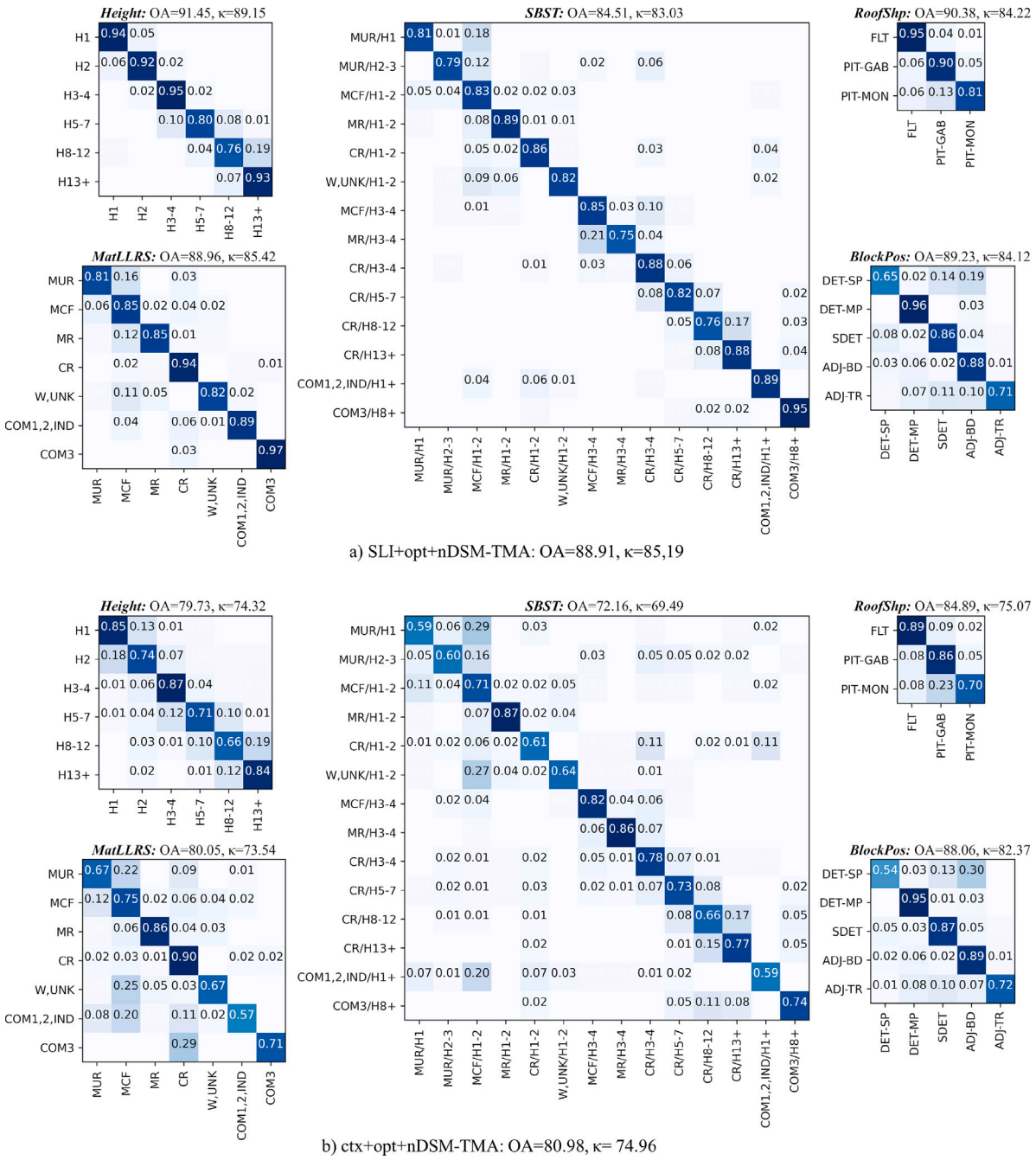

Fig. 12. Confusion matrices, as well task-wise and cross-task mean accuracies (OA and κ in %) of the best M^{3}TC models (y-axis: reference labels, x-axis: predicted labels): (a) *SLI available*; (b) *SLI missing*.

available spatial resolution, and the spatial variability of the intensities of the considered natural hazards (Dabbeek and Silva, 2019). The higher the spatial resolution of the exposure model, the greater its flexibility in meeting these requirements effectively. At the same time, a high thematic resolution is crucial to adequately capture the specific vulnerabilities of buildings to different natural hazards (Pittore et al., 2017; Silva et al., 2022).

5. Summary and conclusion

This paper investigates the integration of heterogeneous multimodal geo-image data (*i.e.*, SLI, VHR optical remote sensing, and nDSM data) for vulnerability-related multicriteria characterization of buildings exposed to natural hazards. It introduces a deep multimodal multitask learning methodology, designed to enable a synergistic integration and efficient classification of these complementary datasets. Herein, *task-wise modality attention* (TMA) fusion is employed to optimize the synergistic utilization of multimodal input data across multiple inference tasks, weighting their feature representations based on the specific requirements of each task. To leverage the highly valuable yet limited semantic information of SLI façade views in a spatially continuous manner, the *SLI spatial context encoder* is proposed. This transformer-based encoder exploits spatial correlations among structural building characteristics to generate meaningful representations as substitutes for data points lacking SLI coverage. The proposed methods facilitate the creation of an area-wide exposure dataset with a unique combination of spatial and thematic resolution, paired with high reliability.

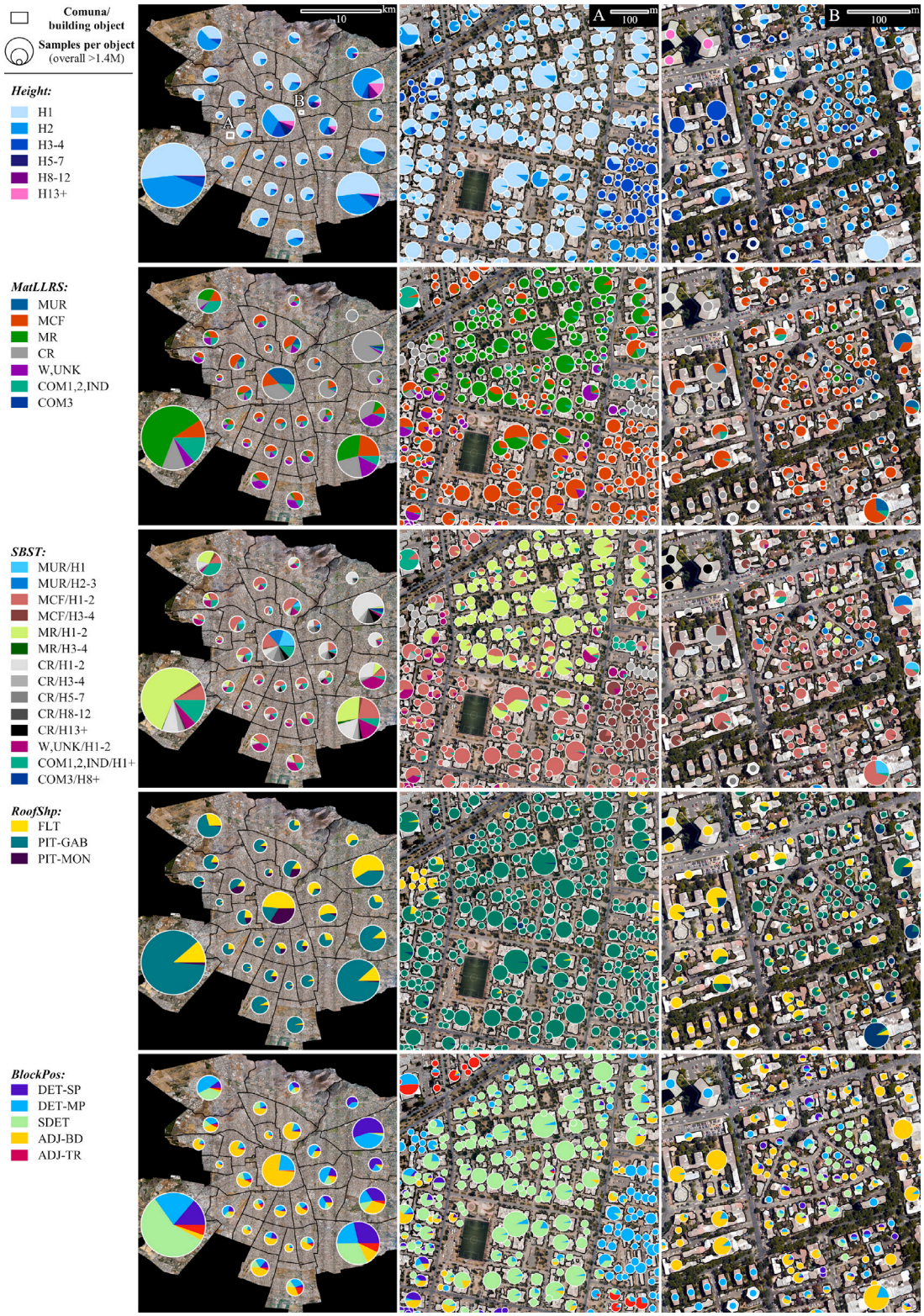

Fig. 13. Exposure model for Santiago de Chile. The rows show the spatial distribution of the five predicted vulnerability-related building characteristics: building height, LLRS material type, SBST, roof shape, and block position. Left column: data points aggregated at the comuna level; center and right columns (zoom boxes A and B, respectively): aggregation at the building object level.

Considering the data scenarios—*SLI available* and *SLI missing*—the experimental evaluations for classifying the five addressed target variables (*height*, *LLRS material*, *SBST*, *roof shape*, and *block position*) demonstrated positive synergies across all input modality combinations, resulting in significant accuracy gains. Accordingly, under both scenarios, the fusion of all considered modalities yields the highest accuracies. The results demonstrate that TMA fusion of modalities consistently outperforms the considered benchmarks, including feature concatenation and decision-level fusion. The highest estimated generalization accuracies are achieved for data points with SLI coverage, with cross-task mean

values reaching up to $OA = 88.91\%$ and $\kappa = 85.19\%$. For datapoints with missing SLI cross-task mean accuracy values of up to $OA = 80.98\%$ and $\kappa = 74.96\%$ were achieved.

In both data scenarios, the integration of SLI-based information proves particularly valuable for achieving accurate and thematically differentiated structural characterization of buildings—either through its direct use in the former case or via SLI spatial context encoding in the latter. This underscores the pivotal importance of the rich semantics in SLI for extracting structural characteristics of exposed buildings relevant to vulnerability assessment, leveraging the geo-image modalities examined in this study.

The findings of this research highlight that integrating ground-based SLI and top-view remote sensing data with tailored DL models is a promising approach for automating the generation of area-wide exposure models with high spatial and thematic resolution—an essential requirement for effective disaster mitigation and management, particularly when considering vulnerability and risk across multiple natural hazards.

CRediT authorship contribution statement

Patrick Aravena Pelizari: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Christian Geiß:** Conceptualization, Investigation, Supervision, Validation, Writing – review & editing. **Hannes Taubenböck:** Conceptualization, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study has been conducted as part of the project RIESGOS 2.0 (03G0905A), funded by the German Federal Ministry of Education and Research (BMBF). We thank Dorothee Stiller for providing the building object and nDSM data.

References

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J., 2010. Google street view: Capturing the world at street level. Computer 43 (6), 32–38. http://dx.doi.org/10.1109/mc.2010.170.

Aravena Pelizari, P., Geiß, C., Aguirre, P., Santa María, H., Merino Peña, Y., Taubenböck, H., 2021. Automated building characterization for seismic risk assessment using street-level imagery and deep learning. ISPRS J. Photogramm. Remote Sens. 180, 370–386. http://dx.doi.org/10.1016/j.isprsjprs.2021.07.004.

Aravena Pelizari, P., Geiß, C., Groth, S., Taubenböck, H., 2023. Deep multitask learning with label interdependency distillation for multicriteria street-level image classification. ISPRS J. Photogramm. Remote Sens. 204, 275–290. http://dx.doi.org/10.1016/j.isprsjprs.2023.09.001.

Aravena Pelizari, P., Spröhnle, K., Geiß, C., Schoepfer, E., Plank, S., Taubenböck, H., 2018. Multi-sensor feature fusion for very high spatial resolution built-up area extraction in temporary settlements. Remote Sens. Environ. 209, 793–807. http://dx.doi.org/10.1016/j.rse.2018.02.025.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. http://dx.doi.org/10.48550/arXiv.1607.06450.

Baltrusaitis, T., Ahuja, C., Morency, L.-P., 2019. Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41 (2), 423–443. http://dx.doi.org/10.1109/tpami.2018.2798607.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: A review. Landsc. Urban Plan. 215, 104217. http://dx.doi.org/10.1016/j.landurbplan.2021.104217.

Calvi, G., Pinho, R., Magenes, G., Bommer, J., Restrepo-Vélez, L., Crowley, H., 2006. Development of seismic vulnerability assessment methodologies over the past 30 years. ISET J. Earthq. Technol. 43 (3), 75–104.

Chen, B., Feng, Q., Niu, B., Yan, F., Gao, B., Yang, J., Gong, J., Liu, J., 2022. Multi-modal fusion of satellite and street-view images for urban village classification based on a dual-branch deep neural network. Int. J. Appl. Earth Obs. Geoinf. 109 (102794), http://dx.doi.org/10.1016/j.jag.2022.102794.

Chen, S., Shi, Y., Xiong, Z., Zhu, X.X., 2023. HTC-DC Net: Monocular height estimation from single remote sensing images. IEEE Trans. Geosci. Remote Sens. 61, 1–18. http://dx.doi.org/10.1109/tgrs.2023.3321255.

Cheng, G., Xie, X., Han, J., Guo, L., Xia, G.-S., 2020. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 13, 3735–3756. http://dx.doi.org/10.1109/JSTARS.2020.3005403.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46. http://dx.doi.org/10.1177/001316446002000104.

Dabbeek, J., Silva, V., 2019. Modeling the residential building stock in the middle east for multi-hazard risk assessment. Nat. Hazards 100 (2), 781–810. http://dx.doi.org/10.1007/s11069-019-03842-7.

d'Angelo, P., Reinartz, P., 2011. Semiglobal matching results on the ISPRS stereo matching benchmark. ISPRS Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XXXVIII-4/W19, 79–84. http://dx.doi.org/10.5194/isprsarchives-XXXVIII-4-W19-79-2011.

Dimitrovski, I., Kitanovski, I., Kocev, D., Simidjievski, N., 2023. Current trends in deep learning for earth observation: An open-source benchmark arena for image classification. ISPRS J. Photogramm. Remote Sens. 197, 18–35. http://dx.doi.org/10.1016/j.isprsjprs.2023.01.014.

Dodman, D., Hayward, B., Pelling, M., Castan Broto, V., Chow, W., Chu, E., Dawson, R., Khirfan, L., McPhearson, T., Prakash, A., Zheng, Y., Ziervogel, G., 2022. Cities, settlements and key infrastructure. In: Pörtner, H.-O., Roberts, D., Tignor, M., Poloczanska, E., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., Rama, B. (Eds.), Climate Change 2022: Impacts, Adaptation and Vulnerability. Cambridge University Press, Cambridge, UK and New York,NY, USA, pp. 907–1040. http://dx.doi.org/10.1017/9781009325844.008.

Douglas, J., 2007. Physical vulnerability modelling in natural hazard risk assessment. Nat. Hazards Earth Syst. Sci. 7 (2), 283–288. http://dx.doi.org/10.5194/nhess-7-283-2007.

Esquivel-Salas, L.C., Schmidt-Díaz, V., Pittore, M., Hidalgo-Leiva, D., Haas, M., Moya-Fernández, A., 2022. Remote structural characterization of thousands of buildings from San Jose, Costa Rica. Front. Built Environ. 8, 947329. http://dx.doi.org/10.3389/fbuil.2022.947329.

Geiß, C., Aravena Pelizari, P., Marconcini, M., Sengara, W., Edwards, M., Lakes, T., Taubenböck, H., 2015. Estimation of seismic building structural types using multi-sensor remote sensing and machine learning techniques. ISPRS J. Photogramm. Remote Sens. 104, 175–188. http://dx.doi.org/10.1016/j.isprsjprs.2014.07.016.

Geiß, C., Priesmeier, P., Aravena Pelizari, P., Soto Calderon, A.R., Schoepfer, E., Riedlinger, T., Villar Vega, M., Santa María, H., Gómez Zapata, J.C., Pittore, M., So, E., Fekete, A., Taubenböck, H., 2023. Benefits of global earth observation missions for disaggregation of exposure data and earthquake loss modeling: evidence from Santiago de Chile. Nat. Hazards 119 (2), 779–804. http://dx.doi.org/10.1007/s11069-022-05672-6.

Geiß, C., Taubenböck, H., 2013. Remote sensing contributing to assess earthquake risk: from a literature review towards a roadmap. Nat. Hazards 68 (1), 7–48. http://dx.doi.org/10.1007/s11069-012-0322-2.

Geiß, C., Thoma, M., Pittore, M., Wieland, M., Dech, S., Taubenböck, H., 2017. Multitask active learning for characterization of built environments with multisensor earth observation data. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 10 (12), 5583–5597. http://dx.doi.org/10.1109/jstars.2017.2748339.

Gill, J.C., Malamud, B.D., 2014. Reviewing and visualizing the interactions of natural hazards. Rev. Geophys. 52 (4), 680–722. http://dx.doi.org/10.1002/2013rg000445.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. In: Proceedings of Machine Learning Research, vol .9, PMLR, pp. 249–256.

Gomez-Chova, L., Tuia, D., Moser, G., Camps-Valls, G., 2015. Multimodal classification of remote sensing images: A review and future directions. Proc. IEEE 103 (9), 1560–1584. http://dx.doi.org/10.1109/jproc.2015.2449668.

Gómez Zapata, J.C., Brinckmann, N., Harig, S., Zafrir, R., Pittore, M., Cotton, F., Babeyko, A., 2021. Variable-resolution building exposure modelling for earthquake and tsunami scenario-based risk assessment: an application case in Lima, Peru. Nat. Hazards Earth Syst. Sci. 21 (11), 3599–3628. http://dx.doi.org/10.5194/nhess-21-3599-2021.

Gonzalez, D., Rueda-Plata, D., Acevedo, A.B., Duque, J.C., Ramos-Pollán, R., Betancourt, A., García, S., 2020. Automatic detection of building typology using deep learning methods on street level images. Build. Environ. 177, 106805. http://dx.doi.org/10.1016/j.buildenv.2020.106805.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision. ICCV, IEEE, http://dx.doi.org/10.1109/iccv.2017.322.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: 2015 IEEE International Conference on Computer Vision. ICCV, pp. 1026–1034. http://dx.doi.org/10.1109/ICCV.2015.123.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Herold, M., Liu, X., Clarke, K.C., 2003. Spatial metrics and image texture for mapping urban land use. Photogramm. Eng. Remote. Sens. 69 (9), 991–1001. http://dx.doi.org/10.14358/pers.69.9.991.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hoffmann, E.J., Abdulahhad, K., Zhu, X.X., 2023. Using social media images for building function classification. Cities 133, 104107. http://dx.doi.org/10.1016/j.cities.2022.104107.

Hoffmann, E.J., Wang, Y., Werner, M., Kang, J., Zhu, X.X., 2019. Model fusion for building type classification from aerial and street view images. Remote. Sens. 11 (11), 1259. http://dx.doi.org/10.3390/rs11111259.

Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B., 2021. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Trans. Geosci. Remote Sens. 59 (5), 4340–4354. http://dx.doi.org/10.1109/tgrs.2020.3016820.

Hosseinpour, H., Samadzadegan, F., Javan, F.D., 2022. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. ISPRS J. Photogramm. Remote Sens. 184, 96–115. http://dx.doi.org/10.1016/j.isprsjprs.2021.12.007.

Hou, Y., Quintana, M., Khomiakov, M., Yap, W., Ouyang, J., Ito, K., Wang, Z., Zhao, T., Biljecki, F., 2024. Global streetscapes — A comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. ISPRS J. Photogramm. Remote Sens. 215, 216–238. http://dx.doi.org/10.1016/j.isprsjprs.2024.06.023.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7132–7141. http://dx.doi.org/10.1109/CVPR.2018.00745.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2261–2269. http://dx.doi.org/10.1109/CVPR.2017.243.

Huang, B., Zhao, B., Song, Y., 2018. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. Remote Sens. Environ. 214, 73–86. http://dx.doi.org/10.1016/j.rse.2018.04.050.

Ibrahim, M.R., Haworth, J., Cheng, T., 2020. Understanding cities with machine eyes: A review of deep computer vision in urban analytics. Cities 96, 102481. http://dx.doi.org/10.1016/j.cities.2019.102481.

IDE, 2015. Fotografía Aérea Del Gran Santiago Año 2014. Infraestructura de Datos Geoespaciales, Chile, Online: https://www.ide.cl/descargas/capas/economia/Fotografia-aerea-Gran-Santiago.rar (Zugriff am 12.04.2021).

Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X.X., 2018. Building instance classification using street view images. ISPRS J. Photogramm. Remote Sens. 145, 44–59. http://dx.doi.org/10.1016/j.isprsjprs.2018.02.006.

Kieu, N., Nguyen, K., Nazib, A., Fernando, T., Fookes, C., Sridharan, S., 2024. Multimodal colearning meets remote sensing: Taxonomy, state of the art, and future works. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17, 7386–7409. http://dx.doi.org/10.1109/jstars.2024.3378348.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. http://dx.doi.org/10.48550/ARXIV.1412.6980, arXiv:1412.6980.

Kraff, N.J., Wurm, M., Taubenböck, H., 2020. Uncertainties of human perception in visual image interpretation in complex urban environments. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 13, 4229–4241. http://dx.doi.org/10.1109/jstars.2020.3011543.

Lee, T., Kashyap, R., Chu, C., 1994. Building skeleton models via 3-D medial surface axis thinning algorithms. CVGIP, Graph. Models Image Process. 56 (6), 462–478. http://dx.doi.org/10.1006/cgip.1994.1042.

Lefevre, S., Tuia, D., Wegner, J.D., Produit, T., Nassar, A.S., 2017. Toward seamless multiview scene analysis from satellite to street level. Proc. IEEE 105 (10), 1884–1899. http://dx.doi.org/10.1109/jproc.2017.2684300.

Li, Z., Chen, B., Wu, S., Su, M., Chen, J.M., Xu, B., 2024b. Deep learning for urban land use category classification: A review and experimental assessment. Remote Sens. Environ. 311, 114290. http://dx.doi.org/10.1016/j.rse.2024.114290.

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J., 2022. Deep learning in multimodal remote sensing data fusion: A comprehensive review. Int. J. Appl. Earth Obs. Geoinf. 112, 102926. http://dx.doi.org/10.1016/j.jag.2022.102926.

Li, Q., Mou, L., Sun, Y., Hua, Y., Shi, Y., Zhu, X.X., 2024a. A review of building extraction from remote sensing imagery: Geometrical structures and semantic attributes. IEEE Trans. Geosci. Remote Sens. 62, 1–15. http://dx.doi.org/10.1109/tgrs.2024.3369723.

Liuzzi, M., Aravena Pelizari, P., Geiß, C., Masi, A., Tramutoli, V., Taubenböck, H., 2019. A transferable remote sensing approach to classify building structural types for seismic risk analyses: the case of Val d'Agri area (Italy). Bull. Earthq. Eng. 17 (9), 4825–4853. http://dx.doi.org/10.1007/s10518-019-00648-7.

Machado, G., Ferreira, E., Nogueira, K., Oliveira, H., Brito, M., Gama, P.H.T., Santos, J.A.d., 2021. Airound and CV-BrCT: Novel multiview datasets for scene classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14, 488–503. http://dx.doi.org/10.1109/jstars.2020.3033424.

Machado, G., Pereira, M.B., Nogueira, K., Santos, J.A.D., 2023. Facing the void: Overcoming missing data in multi-view imagery. IEEE Access 11, 12547–12554. http://dx.doi.org/10.1109/access.2022.3231617.

Martins, V.S., Kaleita, A.L., Gelder, B.K., da Silveira, H.L., Abe, C.A., 2020. Exploring multiscale object-based convolutional neural network (multi-OCNN) for remote sensing image classification at high spatial resolution. ISPRS J. Photogramm. Remote Sens. 168, 56–73. http://dx.doi.org/10.1016/j.isprsjprs.2020.08.004.

Mena, F., Arenas, D., Nuske, M., Dengel, A., 2024. Common practices and taxonomy in deep multiview fusion for remote sensing applications. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17, 4797–4818. http://dx.doi.org/10.1109/jstars.2024.3361556.

Meng, C., Song, Y., Ji, J., Jia, Z., Zhou, Z., Gao, P., Liu, S., 2021. Automatic classification of rural building characteristics using deep learning methods on oblique photography. Build. Simul. 15 (6), 1161–1174. http://dx.doi.org/10.1007/s12273-021-0872-x.

Müller, K., Leppich, R., Geiß, C., Borst, V., Aravena Pelizari, P., Kounev, S., Taubenböck, H., 2023. Deep neural network regression for normalized digital surface model generation with sentinel-2 imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 8508–8519. http://dx.doi.org/10.1109/jstars.2023.3297710.

Mutreja, G., Bittner, K., 2023. Evaluating convnet and transformer based self-supervised algorithms for building roof form classification. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLVIII-1/W2-2023, 315–321. http://dx.doi.org/10.5194/isprs-archives-xlviii-1-w2-2023-315-2023.

Neupane, B., Horanont, T., Aryal, J., 2021. Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. Remote. Sens. 13 (4), http://dx.doi.org/10.3390/rs13040808, 808.

Ogawa, Y., Zhao, C., Oki, T., Chen, S., Sekimoto, Y., 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 J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 1740–1755. http://dx.doi.org/10.1109/jstars.2023.3237509.

Perko, R., Raggam, H., Gutjahr, K.H., Schardt, M., 2015. Advanced DTM generation from very high resolution satellite stereo images. ISPRS Ann. Photogramm. Remote. Sens. Spatial Inf. Sci. II-3/W4, 165–172. http://dx.doi.org/10.5194/isprsannals-ii-3-w4-165-2015.

Pittore, M., Haas, M., Megalooikonomou, K.G., 2018. Risk-oriented, bottom-up modeling of building portfolios with faceted taxonomies. Front. Built Environ. 4, http://dx.doi.org/10.3389/fbuil.2018.00041.

Pittore, M., Wieland, M., Fleming, K., 2017. Perspectives on global dynamic exposure modelling for geo-risk assessment. Nat. Hazards 86 (S1), 7–30. http://dx.doi.org/10.1007/s11069-016-2437-3.

Qiao, Z., Yuan, X., 2021. Urban land-use analysis using proximate sensing imagery: a survey. Int. J. Geogr. Inf. Sci. 35 (11), 2129–2148. http://dx.doi.org/10.1080/13658816.2021.1919682.

Ramachandram, D., Taylor, G.W., 2017. Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Process. Mag. 34 (6), 96–108. http://dx.doi.org/10.1109/msp.2017.2738401.

Rawat, W., Wang, Z., 2017. Deep convolutional neural networks for image classification: A comprehensive review. Neural Comput. 29 (9), 2352–2449. http://dx.doi.org/10.1162/neco_a_00990.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252. http://dx.doi.org/10.1007/s11263-015-0816-y.

Rußwurm, M., Körner, M., 2020. Self-attention for raw optical satellite time series classification. ISPRS J. Photogramm. Remote Sens. 169, 421–435. http://dx.doi.org/10.1016/j.isprsjprs.2020.06.006.

Santa María, H., Hube, M.A., Rivera, F., Yepes-Estrada, C., Valcárcel, J.A., 2017. Development of national and local exposure models of residential structures in Chile. Nat. Hazards 86 (S1), 55–79. http://dx.doi.org/10.1007/s11069-016-2518-3.

Sarabandi, P., Kiremidjian, A., 2007. Development of Algorithms for Building Inventory Compilation Through Remote Sensing and Statistical Inferencing. Technical Report 158, The John A. Blume Earthquake Engineering Center, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA.

Schmitt, M., Zhu, X.X., 2016. Data fusion and remote sensing: An ever-growing relationship. IEEE Geosci. Remote. Sens. Mag. 4 (4), 6–23. http://dx.doi.org/10.1109/mgrs.2016.2561021.

Schuster, M., Paliwal, K., 1997. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45 (11), 2673–2681. http://dx.doi.org/10.1109/78.650093.

Silva, V., Brzev, S., Scawthorn, C., Yepes, C., Dabbeek, J., Crowley, H., 2022. A building classification system for multi-hazard risk assessment. Int. J. Disaster Risk Sci. 13 (2), 161–177. http://dx.doi.org/10.1007/s13753-022-00400-x.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (56), 1929–1958, URL: http://jmlr.org/papers/v15/srivastava14a.html.

Srivastava, S., Vargas-Muñoz, J.E., Tuia, D., 2019. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. Remote Sens. Environ. 228, 129–143. http://dx.doi.org/10.1016/j.rse.2019.04.014.

Stiller, D., Stark, T., Wurm, M., Dech, S., Taubenböck, H., 2019. Large-scale building extraction in very high-resolution aerial imagery using mask R-CNN. In: 2019 Joint Urban Remote Sensing Event. JURSE, pp. 1–4. http://dx.doi.org/10.1109/JURSE.2019.8808977.

Stiller, D., Wurm, M., Stark, T., D'Angelo, P., Stebner, K., Dech, S., Taubenböck, H., 2021. Spatial parameters for transportation: A multi-modal approach for modelling the urban spatial structure using deep learning and remote sensing. J. Transp. Land Use 14 (1), http://dx.doi.org/10.5198/jtlu.2021.1855.

Sun, M., Zhang, F., Duarte, F., Ratti, C., 2022. Understanding architecture age and style through deep learning. Cities 128, 103787. http://dx.doi.org/10.1016/j.cities.2022.103787.

Tan, M., Le, Q., 2021. EfficientNetV2: Smaller models and faster training. In: Proceedings of the 38th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 139, PMLR, pp. 10096–10106.

Taubenböck, H., Mast, J., Geiß, C., Wurm, M., Esch, T., Seto, K., 2024. Global differences in urbanization dynamics from 1985 to 2015 and outlook considering IPCC climate scenarios. Cities 151, 105117. http://dx.doi.org/10.1016/j.cities.2024.105117.

Taubenböck, H., Roth, A., Dech, S., Mehl, H., Münich, J., Stempniewski, L., Zschau, J., 2009. Assessing building vulnerability using synergistically remote sensing and civil engineering. In: Kreck, A., Rumor, M., Zlatanova, S., Fendel, E. (Eds.), Urban and Regional Data Management. Taylor & Francis Group, pp. 287–300.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46, 234. http://dx.doi.org/10.2307/143141.

UNDRR, 2022. Global Assessment Report on Disaster RiskReduction 2022. Our World at Risk: Transforming Governance for a Resilient Future. United Nations Office for Disaster Risk Reduction, Geneva.

UNISDR, 2015. Sendai Framework for Disaster Risk Reduction 2015–2030. United Nations International Strategy for Disaster Reduction, Geneva.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 30, Curran Associates, Inc., URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, J., Zheng, Y., Wang, M., Shen, Q., Huang, J., 2021. Object-scale adaptive convolutional neural networks for high-spatial resolution remote sensing image classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14, 283–299. http://dx.doi.org/10.1109/jstars.2020.3041859.

Wieland, M., Pittore, M., Parolai, S., Zschau, J., Moldobekov, B., Begaliev, U., 2012. Estimating building inventory for rapid seismic vulnerability assessment: Towards an integrated approach based on multi-source imaging. Soil Dyn. Earthq. Eng. 36, 70–83. http://dx.doi.org/10.1016/j.soildyn.2012.01.003.

Wyss, M., Rosset, P., 2013. Mapping seismic risk: the current crisis. Nat. Hazards 68 (1), 49–52. http://dx.doi.org/10.1007/s11069-012-0256-8.

Xie, Y., Tian, J., Zhu, X.X., 2023. A co-learning method to utilize optical images and photogrammetric point clouds for building extraction. Int. J. Appl. Earth Obs. Geoinf. 116, 103165. http://dx.doi.org/10.1016/j.jag.2022.103165.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.-Y., 2020. On layer normalization in the transformer architecture. In: Proceedings of the 37th International Conference on Machine Learning. Vol. 108.

Yu, Q., Wang, C., McKenna, F., Yu, S.X., Taciroglu, E., Cetiner, B., Law, K.H., 2020. Rapid visual screening of soft-story buildings from street view images using deep learning classification. Earthq. Eng. Eng. Vib. 19 (4), 827–838. http://dx.doi.org/10.1007/s11803-020-0598-2.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018. An object-based convolutional neural network (OCNN) for urban land use classification. Remote Sens. Environ. 216, 57–70. http://dx.doi.org/10.1016/j.rse.2018.06.034.

Zhang, L., Wang, G., Sun, W., 2023. Automatic identification of building structure types using unmanned aerial vehicle oblique images and deep learning considering facade prior knowledge. Int. J. Digit. Earth 16 (1), 3348–3367. http://dx.doi.org/10.1080/17538947.2023.2247390.

Zhang, F., Wu, L., Zhu, D., Liu, Y., 2019. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. ISPRS J. Photogramm. Remote Sens. 153, 48–58. http://dx.doi.org/10.1016/j.isprsjprs.2019.04.017.

Zhao, M., Meng, Q., Wang, L., Zhang, L., Hu, X., Shi, W., 2024. Towards robust classification of multi-view remote sensing images with partial data availability. Remote Sens. Environ. 306, 114112. http://dx.doi.org/10.1016/j.rse.2024.114112.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2018. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40 (6), 1452–1464. http://dx.doi.org/10.1109/tpami.2017.2723009.

Zhou, Y., Tan, Y., Wen, Q., Wang, W., Li, L., Li, Z., 2023. Deep multimodal fusion model for building structural type recognition using multisource remote sensing images and building-related knowledge. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 9646–9660. http://dx.doi.org/10.1109/jstars.2023.3323484.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote. Sens. Mag. 5 (4), 8–36. http://dx.doi.org/10.1109/mgrs.2017.2762307.