On the Information Transfer Between Imagery, Point Clouds, and Meshes for Multi-Modal Semantics Utilizing Geospatial Data

Over the past years, geospatial data acquisition has become more redundant, more complete, faster, and denser – both spatially and temporally. Sensors such as cameras and LiDAR scanners facilitate multi-modal capturing of our world via imagery and Point Clouds (PCs). The increasing availability of such data calls for automated multi-modal processing and scene analysis (*multi-modal semantics*). Textured meshes integrate both representations by wiring the PC and texturing the reconstructed surface elements (faces) with high-resolution imagery. Meshes are adaptive to the underlying mapped geometry due to their graph structure composed of non-uniform and non-regular faces. Hence, the mesh is a memory-efficient realistic-looking 3D map of the real world – easily understandable even for non-experts.
For these reasons, we primarily strive for semantic segmentation of meshes, while integrating information from images and the LiDAR sensor. In particular, we head for multi-modal semantics utilizing supervised learning. However, publicly available annotated geospatial mesh data has been rare at the beginning of the thesis. Therefore, annotating mesh data has to be done beforehand. To kill two birds with one stone, we aim for a multi-modal fusion that enables multi-modal enhancement of entity descriptors and semi-automatic data annotation leveraging publicly available annotations of non-mesh data. We propose a novel holistic geometry-driven association mechanism that explicitly integrates entities of the three modalities imagery, PC, and mesh (see Table 1). The mesh is the core modality of the inter-modal association coupling the three modalities by the following bilateral subprocesses:

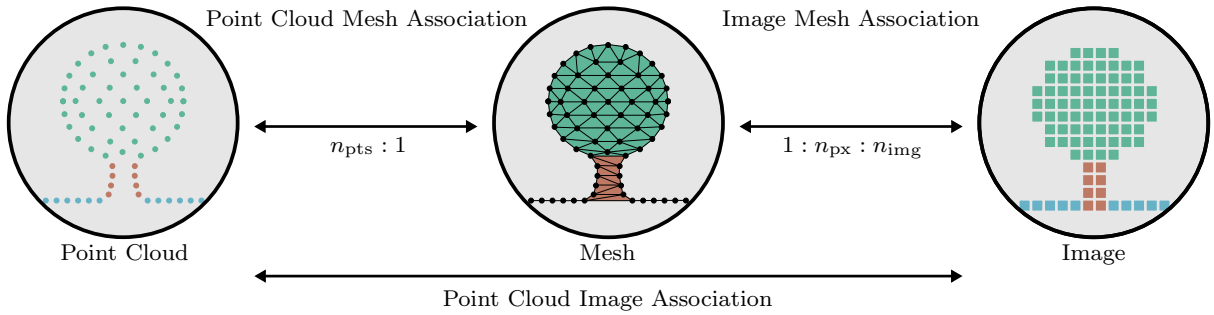
1. Point Cloud Mesh Association (PCMA): Linking each face with several points.
2. Image Mesh Association (ImgMA): Linking each face with several pixels in different images.
3. Point Cloud Image Association (PCImgA): Linking points and pixels in different images while checking the visibility via the mesh (in contrast to a simple, occlusion-agnostic projection). PCImgA is achieved by combining PCMA and ImgMA.

The established entity relationships between pixels, points, and faces enable the *information transfer between imagery, point clouds, and meshes* (see Figure 1) in a two-fold manner: (i) feature transfer (measured or engineered) to enhance modality-specific entities to multi-modal descriptors and (ii) label transfer (predicted or annotated), e.g., to reduce the manual annotation effort by transferring labels to other modalities.

Figure 1: The explicit linking of pixels, points, and faces (multi-modal data fusion, see Table 1) enables the information exchange across the three modalities PC (*left*), mesh (*center*), and imagery (*right*). The figure exemplarily depicts the label propagation from the manually annotated PC to the mesh and an oblique image (for a subset of Hessigheim 3D). Faces that cannot be linked to points remain without a label (depicted in black). Pixels that are linked to an unlabeled face are colored in black. Background and non-associated pixels are colored in reddish-brown.

Taken together, the proposed entity linking serves as an integrative backbone and injects great versatility into the semantic segmentation of *geospatial data* (see Figure 2). Imagery, PCs, and meshes can be semantically segmented with classifiers trained on any of these modalities utilizing features derived from any of these modalities (multi-modality). Particularly, we can semantically segment a modality by training a classifier on the same modality (direct approach) or by transferring predictions from other modalities (indirect approach). Hence, any established well-performing modality-specific classifier can be used for semantic segmentation of these modalities – regardless of whether they follow an end-to-end learning or feature-driven scheme.

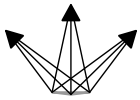
Table 1: Overview of the proposed method that links imagery, point cloud (PC), and mesh via inter-modal subprocesses a) Point Cloud Mesh Association (PCMA), b) Image Mesh Association (ImgMA), and c) Point Cloud Image Association (PCImgA). The top depicts the concept in a pictographic manner. For each association mechanism, the transfer operations depend on the information type (feature or label) and the transfer direction. The pictograms on the right depict the linking of the respective entities.



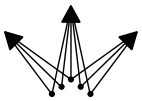
(a) Point Cloud Mesh Association (PCMA)

	Mesh $\mapsto$ PC $(1 : n_{\text{pts}})$	PC $\mapsto$ Mesh $(n_{\text{pts}} : 1)$	
Feature Transfer	Copy Value	Median Aggregation	
Label Transfer	Copy Value	Majority Vote	

(b) Image Mesh Association (ImgMA)

	Mesh $\mapsto$ Img $(1 : n_{\text{px}} : n_{\text{img}})$	Img $\mapsto$ Mesh $(n_{\text{img}} : n_{\text{px}} : 1)$	
Feature Transfer	Copy Value	Median Aggregation	
Label Transfer	Copy Value	Majority Vote	

(c) Point Cloud Image Association (PCImgA)

	PC $\mapsto$ Img	Img $\mapsto$ PC	
Feature Transfer	Copy Value	Median Aggregation	
Label Transfer	Copy Value	Majority Vote	

The implementation follows a tile-wise strategy to facilitate scalability to large-scale geospatial data sets. At the same time, it enables parallel, distributed processing, reducing processing time. The approach is robust against meshing algorithms of different software and can handle 2.5D and 3D meshes. In the light of good scientific practice, we assert that inter-modal discrepancies, co-registration residuals, and the reconstruction quality of the mesh affect the association. We demonstrate the effectiveness of the proposed method on the well-known benchmark data sets Vaihingen 3D and Hessigheim 3D. Figure 3 demonstrates the reliability of the linking and the consistent inter-modal information transfer compared to a simple inter-modal nearest-neighbor interpolation, which cannot cope with data gaps.

Our extensive ablation study reveals the impact of multi-modality for automatic 3D scene interpretation. Figure 4 shows the significant performance increase with increasing multi-modality. The classifier achieves the highest confidence and best performance when both data sources are integrated into the mesh. Whereas LiDAR provides high-quality geometry, imagery provides high-quality textural information.

Since the dawn of the deep learning era, efficient GT generation has become a compelling task. Our linking and transferring methodology enables the consistent labeling of various representations while reducing the manual annotation effort to a single modality (semi-automatic labeling). For instance, labeled 3D data can

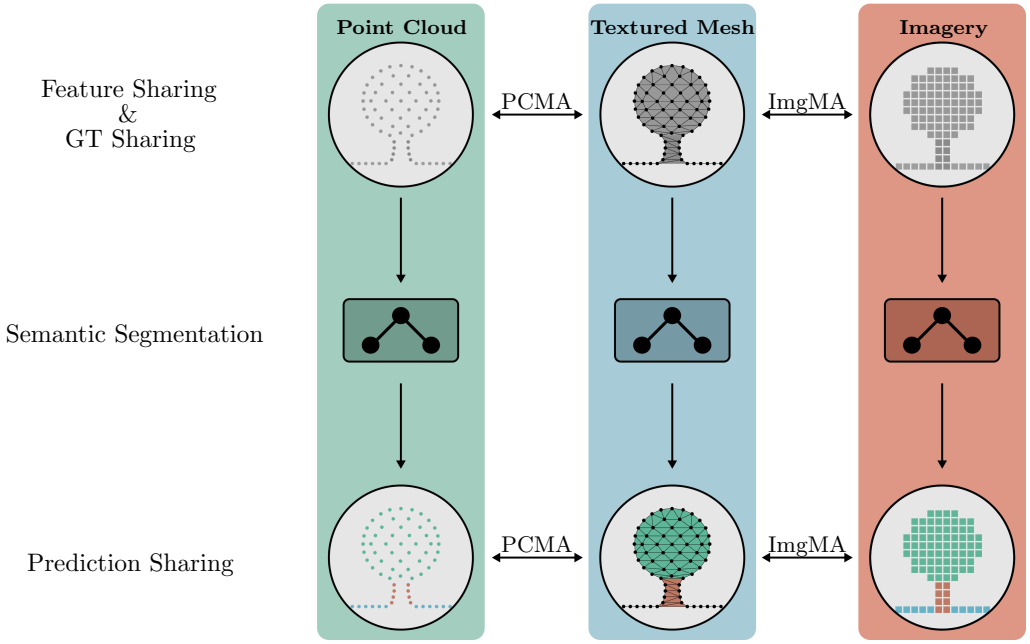


Figure 2: Compact overview of the injected flexibility into the joint semantic segmentation of the modalities imagery (*right*), PC (*left*) and mesh (*center*) utilizing the multi-modal entity linking as the backbone. The multi-modal entity linking enables the flexible sharing of features (inherent or engineered) and labels (Ground Truth (GT) or predictions) – before and after the semantic segmentation with a trained machine learning classifier.

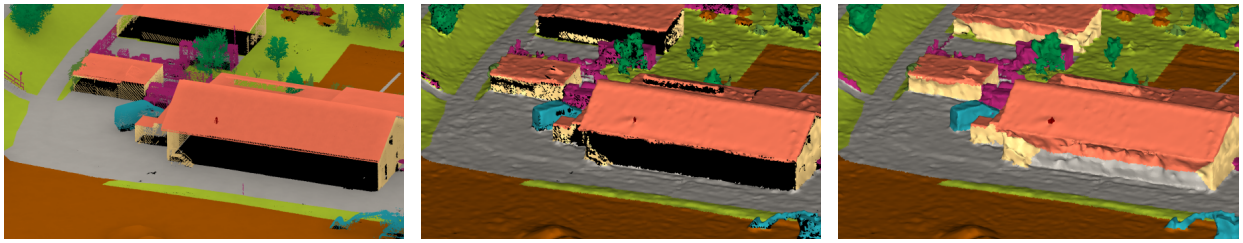
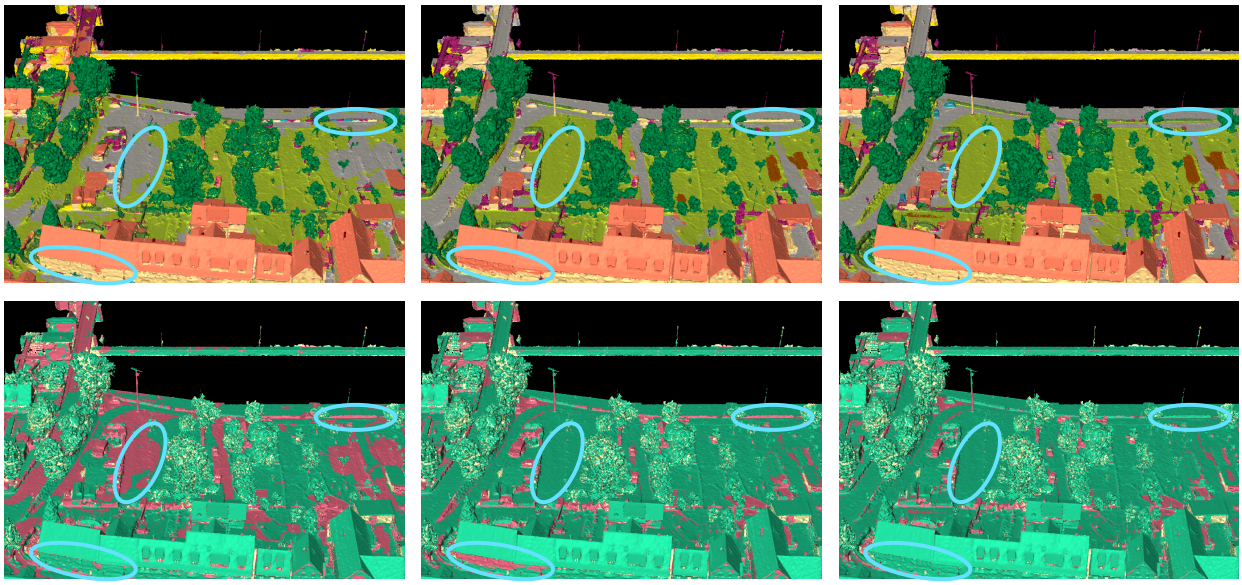


Figure 3: The manual annotations of the Hessigheim 3D LiDAR PC *left* are automatically transferred to the mesh via a) our multi-modal entity linking (*center*) and b) nearest-neighbor interpolation (*right*).

be projected into image space to annotate multiple images at once, avoiding labor-intensive pixel-wise manual annotation. Please note that the multiple epochs of the Hessigheim data consisting of manually annotated PCs and semi-automatically annotated meshes are a result of this thesis and provided to the community as part of the Hessigheim 3D benchmark.

To further reduce the labeling effort to only a few instances on a single modality, we combine the proposed information transfer with so-called Active Learning (AL). The key of AL is to label a sparse, informative subset instead of the entire data set. Additionally, we recruit non-experts for the tedious labeling task. Specifically, annotation time and costs drop from several months and thousands of dollars to a few days and hundreds of dollars when comparing the full annotation by experts with sparse AL-steered labeling done by a crowd of non-experts. Classifiers trained on such sparsely annotated GT perform only 3 pp worse (for $mF1$ and OA) than their passive learning equivalents using 400 times more training points. In our experiments, we note that visualizing the mesh instead of PC improves the crowd's labeling accuracy by up to 3 pp. In turn, the trained classifiers perform significantly better, too. We argue that the realistic-looking appearance of meshes is easier to understand for non-experts than PCs and hence offers better annotation quality.

In summary, we developed a powerful integrative backbone that explicitly links the available data sources, boosting GT generation, multi-modal learning, and joint semantic segmentation of imagery, PCs, and meshes. Besides, we accentuated the mesh and its utility for visualizational purposes and multi-modal semantics. Generally, the proposed methodology captivates with its simplicity and genericity. The method benefits from the recent hybridization trend and advances in data acquisition, adjustment, automatic surface reconstruction, and classifier design, making it very powerful for future developments. Our contribution may further increase the utility and the acceptance of the mesh in photogrammetry and remote sensing.



Mesh Geometry	Mesh Geometry + Texture	Mesh Geometry + Texture + LiDAR
$mF1 = 51.99\,\%$	$mF1 = 60.86\,\%$	$mF1 = 70.09\,\%$
$OA = 68.67\,\%$	$OA = 80.51\,\%$	$OA = 84.84\,\%$

Figure 4: Per-face predictions (*top*) and difference plots (*bottom*) achieved by utilizing a varying amount of the available sensor data: a) pure mesh geometry (*left*), b) textured mesh, i.e., mesh geometry and image content (*center*), and c) textured mesh enhanced by LiDAR attributes (*right*). The snapshots show a close-up of the lock area of the Hessigheim 3D mesh. Correct predictions are shown in *green*; false predictions in *red*. Faces with unknown ground truth are marked in *yellow*.