The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-M-6-2025
ISPRS, EARSeL & DGPF Joint Istanbul Workshop "Topographic Mapping from Space" dedicated to Dr. Karsten Jacobsen's 80th Birthday
29–31 January 2025, Istanbul, Türkiye

# On the Perspectives of Image-to-Lidar Constraints in Dynamic Network Optimisation

Kyriaki Mouzakidou, Thibaut Stoltz, Laurent V. Jospin, Davide A. Cucci, Jan Skaloud

Environmental Sensing Observatory (ESO), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland –
(kyriaki.mouzakidou, laurent.jospin, jan.skaloud)@epfl.ch, thibaut.stoltz@alumni.epfl.ch, davide.cucci@pix4d.com

**Keywords:** Pixel-to-point correspondences, Learned cross-domain matching, Trajectory correction, Dynamic networks.

## Abstract

The evolution of airborne mapping witnesses the introduction of hybrid lidar-camera systems to enhance data collection, i.e. to obtain simultaneously high-density point-cloud and texture. Yet, the common adjustment of both optical data streams is a non-trivial problem due to challenges associated with the different influences of errors affecting their mapping accuracy including those coming from navigation sensors. Stemming from a special form of graph-based optimization, the dynamic networks allow rigorous modeling of spatio-temporal constraints and thus provide the common framework for optimizing original observations from inertial systems with those coming from optical sensors. In this work, we propose a cross-domain observation model that leverages pixel-to-point correspondences as links between imagery and lidar returns. First, we describe how the existence of such correspondences can be introduced into optimizations. Second, we employ a reference dataset to emulate a set of precise pixel-to-point correspondences to assess its prospective impact on the common (rather than cascade) optimization. We report the improvement in the estimated trajectory attitude error with lower quality IMU and thus the point-cloud registration. Finally, we study whether such correspondences could be contained from freely available deep learning networks with the desired accuracy and quality. We conclude that although the introduction of such camera-to-lidar constraints has significant potential, none of the studied machine learning networks can fulfill the requirement on correspondence detection in terms of quality.

## 1. Introduction

Fusing images (2D domain) with point-clouds (3D domain) has become beneficial in aerial applications, e.g. in monitoring and change detection, due to the complementarity of these two modalities (Pöppl et al., 2023). For precise 3D-reconstruction and modeling, aerial photogrammetry is susceptible to vertical offsets, while Airborne Laser Scanning (ALS) is weaker in the horizontal direction as a function of altitude due to direct georeferencing (Glennie, 2007). The evolution of airborne mapping witnesses the introduction of hybrid high-quality camera-laser systems to enhance data collection either closer to the ground with drones (Vallet et al., 2020) or from higher altitudes with aircrafts (CityMapper - Leica Geosystems, 2024), for an accurate, high-density and textured final point-cloud.

### 1.1 Image-Lidar fusion

The common adjustment of the two datasets (2D and 3D) is an active research area due to partially the same input of navigation sensors, yet traditionally separate optimizations. For instance, many challenges associated with the inertial sensors onboard, such as the flight geometry configuration, (e.g. corridor mapping) can impact the co-registration of lidar and images. In this context, the use of spatio-temporal constraints in a common optimization has proven to reduce the influences of these error sources, as shown e.g. in (Cucci et al., 2017) with image-to-image (single-domain, 2D), in (Brun et al., 2022) and (Pöppl et al., 2024) with lidar-to-lidar (single-domain, 3D) and in (Mouzakidou et al., 2024) with both 2D and 3D single-domain constraints, alongside raw inertial and GNSS data in the Dynamic Network (DN) adjustment (Colomina et al., 2004, Cucci and Skaloud, 2019). However, these approaches rely exclusively on single-domain optical constraints either separately or combined.

The continuous development of deep learning has shown its potential to extract cross-domain image-to-lidar constraints creating links between pixels on the imagery and 3D points on the point-cloud. This approach may unlock the new potential of fusing active and passive optical sensors via direct pixel-to-point correspondences, as opposed to the current cascade fusion based on intermediate products (Glira et al., 2019, Hussnain et al., 2021), e.g. point-clouds from dense image matching to lidar, which is sub-optimal unless all correlations are correctly considered. The current deep learning architectures that extract pixel-to-point correspondences are primarily designed for registration tasks (Feng et al., 2019, Pham et al., 2020, Ren et al., 2023, Yao et al., 2024) with certain geographic localization nodes (Li et al., 2023a) and they are mainly tested on terrestrial datasets. More details on this aspect are presented in Sec. 3. We thus observe a research gap in techniques that leverage the extracted 2D-3D correspondences for trajectory optimization, which constitutes the key motivation of this research.

### 1.2 Contributions

In this work, we first describe how a new cross-domain, i.e. pixel-to-point, observation model can be used as spatial constraint in a graph-based optimization algorithm, as described in (Colomina et al., 2004), (Cucci et al., 2017), (Brun et al., 2022) or (Pöppl et al., 2024) together with raw inertial and GNSS observations. In Sec. 2, we demonstrate how this constraint links the original optical sensor observations, i.e. pixels on the image plane with laser vectors from the lidar sensor. In Sec. 3, we revisit the current state-of-the-art (SOTA) deep learning architectures for extracting direct pixel-to-point correspondences and categorize them per type of approach and architecture used. Using emulated pixel-to-point correspondences (Sec. 4.2), we assess the impact of the new type of optical constraints on trajectory determination and point-cloud geo-referencing (Sec. 5.1). Finally, we evaluate on a controlled aerial dataset the SOTA

deep learning architectures that are open source in terms of code and pre-trained network weights (Sec. 5.2). The availability of a ground-truth dataset with centimeter level accuracy allows us to quantify the reliability of the extracted correspondences.

## 2. Pixel-to-point constraint in sensor fusion

This section describes the observation model (spatial constraint between a pixel on the imagery and a 3D point on the point-cloud) and its integration into the existing sensor fusion workflow. The model presumes that pixel-to-point correspondences can be extracted with some uncertainty to constrain trajectory poses (translation and rotation) and system parameters. Subsequently, it is introduced as a spatial constraint in DN.

### 2.1 Observation model

Let us assume that a 2D-3D correspondence can be somewhat established (e.g. via one of the techniques described in Sec. 3) between a 2D pixel $p$ and a 3D lidar point $P$. The distorted image coordinates of pixel $p$ in the *image* frame $c_i$ (here using the symbol *c-camera* for simplification) captured at time $t_i$ are $\ell^{c_i}_{(p)} = [x_d, y_d]^T$, in pixel units. The lidar point $P$ acquired at time $t_j$ from the lidar sensors is projected on an image and thus expressed in the *camera* frame $c_i$ as $p^{c_i}_{(l_j)}$. The notation $l_j$ corresponds to the lidar sensor pose at the given timestamp. Given that $p$ and $P$ are homologous points, we can formulate the following condition in the 2D space (Fig. 1 - element 1):

$$\ell^{c_i}_{(p)} - p^{c_i}_{(l_j)} = 0 + v_{cl} \tag{1}$$

where $v_{cl}$ is a $[2 \times 1]$ vector of zero mean Gaussian noise representing the re-projection residuals in pixels. In the following, we replace 0 with $\ell_{cl}$ since it comprises a so-called *zero*-observation edge, or a *pseudo*-observation in the network terminology, corresponding to no actual sensor measurement. We also swap two sides of Eq. 1 to be consistent with the other DN-observation models summarized in (Mouzakidou et al., 2024).

Considering the collinearity condition for the reprojected lidar point $p^{c_i}_{(l_j)}$ and using the homogeneous coordinates formalism, Eq. 1 can be written as (Fig. 1 - element 2):

$$\ell_{cl} + v_{cl} = \ell^{c_i}_{(p)} - \Xi\left[\pi\left[K\widetilde{\Pi}\left[\widetilde{\Gamma}^m_{b(t_i)}\Gamma^b_c\right]^T P^m_{(l_j)}\right]\right] \tag{2}$$

where, function $\Xi(\cdot)$ models the lens distortions coefficients, that relate distorted and undistorted image coordinates, e.g. the Brown-Conrady distortion model (Brown, 1971), $\pi(\cdot)$ is the projection function[1], $K$ represents the $[3 \times 3]$ camera matrix[2] and $\widetilde{\Pi}$ is an auxiliary matrix that handles the homogeneous coordinates[3]. Term $\widetilde{\Gamma}^m_{b(t_i)}$[4] refers to the pose of the *body* ($b$) frame (usually represented by the internal axes of the inertial system) in the *mapping* ($m$) frame at timestamp $t_i$, while $\Gamma^b_c$ corresponds to the mounting matrix[5], i.e. boresight matrix and lever-arm, of the camera sensor ($c$) in the *body* frame. Finally, $P^m_{(l_j)} = [X^m, Y^m, Z^m, 1]^T$ is the *mapping* frame 3D coordinates in the homogeneous formalism of a point captured by the lidar sensor ($l$) at time $t_j$.

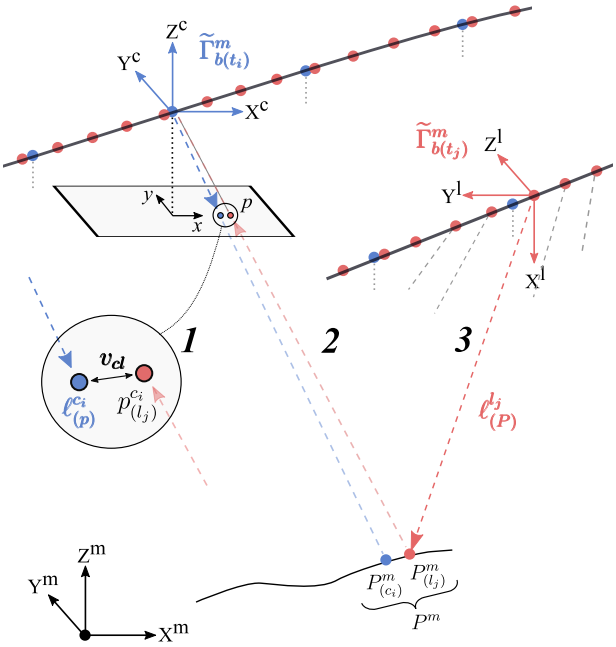


Figure 1. Visual representation of one pixel-to-point correspondence used to constrain two trajectory poses. With blue, we indicate the image poses and observations, while with red the lidar poses and observations.

Point $P^m_{(l_j)}$ needs to be traced back to the original lidar measurement (Fig. 1 - element 3) and expressed as a function of the corresponding 3D lidar vector as $P^m_{(l_j)} = \widetilde{\Gamma}^m_{b(t_j)}\Gamma^b_l \ell^{l_j}_{(P)}$, where $\widetilde{\Gamma}^m_{b(t_j)}$[4] and $\Gamma^b_l$[5] were described before and $\ell^{l_j}_{(P)}$ is the laser vector of point $P$ captured from the lidar pose $l_j$, expressed in the lidar frame. Introducing this relation in Eq. 2 results in Eq. 3 which represents the spatial constraint in DN that links a 2D image pixel to its homologous 3D lidar point observation, stochastically conditioning the trajectory solution (Fig. 1).

$$\ell_{cl} + v_{cl} = \underbrace{\ell^{c_i}_{(p)}}_{\substack{\text{image}\\ \text{observation}}} - \underbrace{\Xi\left[\pi\left[K\widetilde{\Pi}\left[\widetilde{\Gamma}^m_{b(t_i)}\Gamma^b_c\right]^T \widetilde{\Gamma}^m_{b(t_j)}\Gamma^b_l \ell^{l_j}_{(P)}\right]\right]}_{\substack{\text{lidar observation}\\ \text{projected to image coordinates}}} \tag{3}$$

### 2.2 Dynamic network structure

The proposed model (Eq. 3) is introduced as a spatial constraint in the adjustment, complementary to other spatio-temporal constraints summarized in (Mouzakidou et al., 2024), Tab. 1. Following the factor graph formulation of DN described in (Cucci

[1] $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2 : \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \rightarrow \frac{1}{Z}\begin{bmatrix} X \\ Y \end{bmatrix}$

[2] Given the principal distance $c$ and the principal point $[ppx, ppy]$ expressed in pixels, $K = \begin{bmatrix} c & 0 & ppx \\ 0 & c & ppy \\ 0 & 0 & 1 \end{bmatrix}$.

[3] $\widetilde{\Pi} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

[4] Given the $[3 \times 3]$ attitude matrix $R^m_{b(t_i)}$ and the $[3 \times 1]$ position vector $T^m_{b(t_i)}$ of the *body* ($b$) frame in the *mapping* ($m$) frame at time $t_i$, $\widetilde{\Gamma}^m_{b(t_i)} = \begin{bmatrix} R^m_{b(t_i)} & T^m_{b(t_i)} \\ \mathbf{0} & 1 \end{bmatrix}$.

[5] For a given sensor $s$, i.e. camera ($c$) or lidar ($l$), mounting information that comprises the $[3\times3]$ boresight matrix $R^b_s$ and the $[3\times1]$ lever-arm vector $\alpha^b_s$ of the sensor ($s$) in the *body* ($b$) frame are summarized in the matrix $\Gamma^b_s = \begin{bmatrix} R^b_s & -\alpha^b_s \\ \mathbf{0} & 1 \end{bmatrix}$.

et al., 2017), the camera-to-lidar edges connect two pose nodes and optionally, the camera and lidar boresight nodes. Hence, the unknown parameters that are constrained are (i) camera pose at which the image is taken ($\widetilde{\Gamma}^m_{b(t_i)}$), (ii) lidar pose at which the laser vector is acquired ($\widetilde{\Gamma}^m_{b(t_j)}$), (iii) camera internal parameters ($K$ and $\Xi(\cdot)$), (iv) camera and (v) lidar mounting parameters ($\Gamma^b_c$ and $\Gamma^b_l$ respectively), i.e. boresight matrices and lever arms.

## 3. Direct pixel-to-point matching

Here we review the current SOTA "machine-learned" techniques for extracting image-to-lidar correspondences. We can distinguish between two main categories of deep learning architectures serving this purpose: (i) detect-then-match and (ii) detection-free architectures. We summarize them in Tab. 1 and categorize them per type of approach and architecture used.

### 3.1 Detect-then-match networks

In these networks, 2D and 3D keypoints are first detected independently in the image and the point-cloud respectively, and then matched based on their associated descriptors. The workflow described in Fig. 2, consists of the following steps:

1. 2D and 3D keypoints detection in the image and in the point-cloud
2. 2D and 3D patches extraction around each keypoint
3. 2D and 3D feature descriptors extraction
4. Matching pairs of corresponding 2D and 3D features
5. Filtering of the matched 2D-3D pairs.

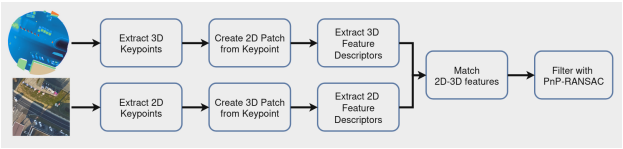


Figure 2. 2D-3D keypoint matching processing workflow

All networks in this category have two separate branches to treat 2D and 3D features separately and then jointly learn the description or matching step of the workflow. From these, only LCD is evaluated later (see Sec. 4.3 and 5.2) as it is the only one provided with open code and weights.

- 2D3D-MatchNet (Feng et al., 2019) is trained and evaluated on an outdoor street dataset. It uses SIFT (Lowe, 2004) and ISS (Zhong, 2009) to extract 2D and 3D features respectively. The network then learns descriptors for 2D and 3D keypoints such that matching pairs have minimal descriptor-space distance while non-matching pairs have maximal distance. Its architecture is based on a 2D CNN (Convolutional Neural Network) and PointNet (Charles et al., 2017) to learn the 2D and 3D features respectively.
- LCD (Pham et al., 2020) is trained and evaluated on two indoor RGB-D datasets, which are used to extract the ground-truth correspondences. It does not depend on an explicit feature extraction method, but implies already extracted correspondences. It employs a dual auto-encoder architecture based on a 2D CNN and PointNet to learn robust feature representations in a common latent space.
- Siam2D3D-Net (Liu et al., 2020) is trained and evaluated on an outdoor street dataset. Similarly to 2D3D-MatchNet, it uses SIFT and ISS to extract 2D and 3D features. Its architecture is similar to 2D3D-MatchNet with the extra step of a spatial transformer network to learn the image feature representations.
- Desc-Matcher (Nadeem et al., 2023) is trained and evaluated on both outdoor street and indoor datasets. As opposed to the three previous networks, it does not require 2D and 3D descriptors to be in a common learned latent space, but rather learns the descriptor matching step of the workflow. As an example, the authors use SIFT for 2D descriptors extraction and 3D-SIFT (Rusu and Cousins, 2011) and RIFT (Lazebnik et al., 2005) for 3D descriptors extraction. This approach is similar to SuperGlue (Sarlin et al., 2020) for 2D feature matching.

### 3.2 Detection-free networks

These more recent architectures adopt an end-to-end approach, i.e. the final outputs are the registered datasets where the pixel-to-point correspondences are only intermediate results that are jointly optimized with other parameters. These architectures rely on a *coarse-to-fine* methodology: they first establish *coarse* correspondences at the level of image or point-cloud tiles and then perform *fine* matching of pixels and points. The advantage of these methods over detection-based approaches is the possibility to exploit global contextual information at the patch level. However, they cannot handle local deformations in the point-cloud, e.g. as those caused due to direct orientation with a low-quality trajectory. In the following, we revisit the available detection-free networks, out of which D-GLSNet and VP2P-Match will be later evaluated given the availability of open code and weights.

- P2-Net (Wang et al., 2021) is trained and evaluated on indoor datasets. It employs a dual fully convolutional architecture to map 2D and 3D inputs into a shared latent space. The network is jointly optimized with a descriptor and a detector loss enforcing the similarity of corresponding representations as well as encouraging higher detection scores for discriminative correspondences.
- D-GLSNet (Li et al., 2023a) matches outdoor street lidar point clouds with satellite images. The processing methodology consists of a feature extraction stage using Feature Pyramid Networks for images and KPConv for point clouds, followed by a Transformer-based module for coarse and fine feature matching. A dual-softmax operation is employed to handle many-to-one correspondences due to differing resolutions.
- CorrI2P (Ren et al., 2023) is trained and evaluated on outdoor street datasets. It employs ResNet and SO-Net architectures to embed the image and point-cloud into high-dimensional feature spaces, generating pixel-wise and point-wise features respectively. A symmetric cross-attention fusion module is introduced to detect overlapping regions by mapping features between the 2D and 3D domains.
- 2D3D-MATR (Li et al., 2023b) is trained and evaluated on indoor datasets. It utilizes a transformer-based module to learn global contextual constraints and cross-modality

| Model name | Type of approach | Type of architecture | Training /testing data | Open source availability |
|---|---|---|---|---|
| 2D3D-MatchNet (2019) | detect-then-match | 2D CNN + 3D PointNet | outdoor street | none |
| LCD (2019) | detect-then-match | 2D CNN + 3D PointNet (autoencoders) | indoor | code + weights |
| Desc-Matcher (2023) | detect-then-match | Decision-Tree based Matcher | indoor | code + weights |
| Siam-2D3D-Net (2023) | detect-then-match | 2D CNN + STN + 3D PointNet | outdoor + indoor | none |
| P2-Net (2020) | detection-free | 2D CNN + 3D CNN + Keypoint Detection | indoor | code |
| CorrI2P (2023) | detection-free | 2D ResNet + 3D SO-Net + Overlapping Region Detection | street | none |
| D-GLSNet (2023) | detection-free | 2D pyramid + 3D KPConv + Transformer-Based Matching | satellite images + street point cloud | code |
| 2D3D-MATR (2023) | detection-free | 2D pyramid + 3D KPConv + Multi-Scale Matching | indoor | none |
| VP2P-Match (2023) | detection-free | 2D CNN + 3D [voxel CNN + point CNN] + Intersection Detection + Distance Based Matching | street | code + weights |
| CFI2P (2024) | detection-free | 2D ResNet + 3D PointNet + Hybrid Transformer + Optimal Transport Matching | outdoor street | none |

Table 1. State-of-the-art pixel-to-point matching architectures.

correlations. To address the scale ambiguity caused by perspective projection, a multi-scale patch matching strategy is implemented. This approach constructs a multi-scale pyramid for image patches, allowing the network to find the best matching patches at appropriate resolution levels. However, challenges such as precise fine-level matching and handling complex scenes remain.

- VP2P-Match (Zhou et al., 2023) is trained and evaluated on outdoor street datasets. It consists of a voxel and a pixel CNN branch, as well as complementary point branch to capture spatial patterns and regain lost 3D details during voxelization. It also uses a differentiable probabilistic Perspective-n-Point (PnP) solver to learn a cross-modality latent space to represent pixel features and 3D features by imposing supervision directly on the predicted pose distribution.
- CFI2P (Yao et al., 2024) is trained and evaluated on outdoor street datasets. It leverages a hybrid transformer architecture to enhance image-to-point cloud registration by integrating quantity-aware correspondences between point and pixel patches. This method begins with the extraction of local proxies from image patches and point patches, capturing both global and cross-modal contexts using self-attention and cross-attention mechanisms.

## 4. Experimental evaluation

In this section, we first describe the aerial dataset used for the investigations (Sec. 4.1) and the extraction of the emulated 2D-3D correspondences (Sec. 4.2). Then, we refer to the data preparation (Sec. 4.3) to adapt our airborne dataset to the specifications of each tested pre-trained network. Finally, we present the optimization study cases (Sec. 4.4) to assess the proposed DN observation model.

### 4.1 Dataset

We will employ the controlled ALS dataset presented in (Vallet et al., 2020) to evaluate the performance of the proposed DN observation model and the pixel-to-point matching networks. It is acquired by a helicopter carrying optical and navigation sensors of high and lower accuracy. We focus mainly on inertial sensors and their influence on orientation and mapping performance. The considered images come from an IXAR180 (PhaseOne) with 80 megapixels and a 42 mm lens with pre-calibrated interior orientation. The imagery consists of 87 images with an average Ground Sampling Distance ($GSD_i$) of 3 cm/pix. The point-cloud data has been captured by a medium-range VQ480 (Riegl) lidar, with a nominal density of 70 pts/m² that results in a point-cloud $GSD_l$ between 10 to 20 cm. The three point-clouds obtained from the three flight lines are merged into a single point-cloud counting $\approx 35 \cdot 10^6$ points.

Both optical datasets are geo-referenced with the onboard navigation-grade AIRINS (iXblue), that we consider for the ground-truth datasets. The attitude errors of the optimal recursive smoothing are smaller than $< 0.003°$ (Vallet et al., 2020), i.e. $\approx 1.5$ cm at 250 m ranges (mean flight height), so about $10\times$ smaller than $GSD_l$, and $\approx 0.5$ pix given the $GSD_i$. Thus, we can safely consider this trajectory to create the ground-truth geo-referenced dataset.

### 4.2 Emulated correspondences

To investigate the performance of the proposed DN observation model, we employ the scenario of real ALS data (Sec. 4.1) but first with emulated pixel-to-point correspondences. For their creation, we follow the steps described below. We compute $p^{c_i}_{(l_j)}$ from the laser vector $P^{l_j}_{(l_j)}$ given the reference trajectory and calibrated sensor information, making the correspondences the ground-truth. In Fig. 1, this would mean that the two dots on the image plane almost coincide (up to numerical precision) and that $v_{cl} \approx 0$. More specifically, we follow the steps:

1. Create 3D points in the mapping frame, from the laser vectors in the lidar frame, given the reference trajectory and the calibrated lidar mounting parameters.

2. Project these 3D points on the imagery using the collinearity equation (second part of Eq. 3) and given the reference trajectory, the camera mounting parameters and the calibrated camera internal orientation.

3. Relax the constraints on the generated 2D-3D correspondences ($1\sigma$) to half a pixel in the optimization process.

### 4.3 Correspondences preparation for existing networks evaluation

From the available direct pixel-to-point matching networks (Sec. 3.1 and 3.2), we consider the ones with open code and open learned weights, i.e. LCD, D-GLSNet and VP2P-Macth, to be evaluated on the controlled ALS dataset. Each network has different input requirements, i.e. image and point-cloud patches or tiles, so we prepare the data accordingly as summarized in Fig. 3.

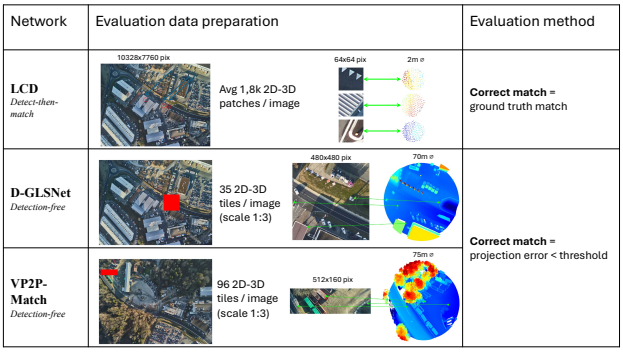


Figure 3. Treatment of the dataset to match the input characteristics of each tested network.

LCD expects 2D patches of $64 \times 64$ pixels and 3D point patches of 1024 *colorized* points. The point-cloud was colorized in the Agisoft Metashape photogrammetric software using the reference trajectory and available imagery. Each image was split into a square grid to create non-overlapping 2D patches. All points in the point-cloud were then projected in the image frame. The center of each 3D patch is selected as the point whose projection is the closest to the center of a 2D patch. Finally, the 3D patch is extracted by selecting all the points around its center, in a sphere of diameter $d = GSD_i \cdot 64 = 0.03\ m \cdot 64 = 1.92\ m$. This resulted in $\approx$ 1800 2D-3D pairs per image. During the evaluation, we consider a match as correct when it corresponds to a ground-truth (reference) pair.

D-GLSNet expects (satellite) images of $480 \times 480$ pixels and their corresponding 3D point-clouds. Since the studied imagery is of much higher resolution, we downsampled the imagery by a factor of 3, i.e. $GSD_i^{'} = factor \cdot GSD_i = 3 \cdot 0.03 = 9\ cm/pix$, and sampled tiles of $480 \times 480$ pixels in the downsampled images to be used as the input resulting in $\approx$ 35 tiles per image. This is a fair compromise between having informative tiles in terms of texture and geometry and maintaining the high resolution of the original imagery. For the 3D tiles extraction, we keep the points that project in each image tile; the center of a 3D tile is randomly chosen among all the points that are projected in a certain radius from the respective 2D center. This randomness in the 3D center selection simulates the imprecision in the prior knowledge of the pose, in which D-GLSNet is robust based on the authors. The radius is set to 5 meters. The 3D tile is formed by the points inside the vertical cylinder (along $Z$ axis) of infinite height, passing through the 3D tile center, and a radius $r_{3Dtile} = GSD_i^{'} \cdot (\frac{480}{\sqrt{2}}) + 5$ (in meters). In that way, we ensure that the 2D tile is always inside the projection of the 3D tile. During the evaluation, a correct match is defined by the re-projection error that should be below a certain threshold, i.e. $2\ m$ given the GSD in our case. It is worth mentioning that due to the voxelization of the whole point-cloud for the treatment and search of points, the output matches do not contain ground-truth lidar points (that are linked to their respective laser vectors), but new points across the voxel grid.

VP2P expects rectangular images of $512 \times 160$ pixels and their corresponding 3D point-clouds. Similarly to D-GLSNet, we use the image downsampling factor of 3 and follow the same procedure to extract image tiles and their respective 3D tiles, resulting in 96 tiles per image. During the evaluation, a correct match is again defined by the re-projection error that should be $2\ m$ given the GSD in our case. It is worth mentioning that the default output of the network downsamples the output image coordinates by a factor of 4. So instead of $512 \times 160$ the output coordinates have a range of $128 \times 40$. This peculiarity of the network brings a certain loss of resolution which as will be shown later can be detrimental.

### 4.4 Optimization cases

Based on the type of spatial constraints used together with the GNSS and raw inertial observations in DN, we study four trajectory determination cases that we compare with the reference:

**Reference trajectory:** Trajectory generated with the navigation-grade IMU which has high geo-referencing accuracy (Sec 4.1) with attitude errors smaller than $< 0.003^{\circ}$ (Vallet et al., 2020). Through that, we obtain by direct geo-referencing the reference point-cloud using the formerly calibrated lidar boresight.

**Case A - [IMU + GNSS] :** Trajectory computed via the loosely coupled integration of IMU readings with the GNSS position solution in a recursive smoother using the software Posproc (Applanix) with an internally designed model for the low-cost IMU. It is used here as a baseline for comparison, to show the impact of the DN adjustment with spatio-temporal constraints.

**Case B - [IMU + GNSS] + 2D-2D + 3D-3D:** DN computed trajectory integrating GNSS and raw inertial readings together with single-domain correspondences, i.e. image-to-image (2D) and lidar-to-lidar (3D). This approach is extensively studied in (Mouzakidou et al., 2022) and (Mouzakidou et al., 2024), and is used here to compare with the proposed approach of using cross-domain spatial constraints.

**Case C - [IMU + GNSS] + 2D-3D:** DN computed trajectory following the proposed approach of integrating GNSS and raw inertial readings with cross-domain correspondences only, i.e. the newly introduced pixel-to-point.

**Case D - [IMU + GNSS] + 2D-2D + 3D-3D + 2D-3D:** DN computed trajectory following the proposed approach of integrating GNSS and raw inertial readings with all available spatial constraints, i.e. cross-domain (pixel-to-point) and single-domain (image-to-image (2D) and lidar-to-lidar (3D)) correspondences.

## 5. Results & analysis

The results are split into two categories assessing: (i) the impact of the proposed DN observation model given the emulated 2D-3D correspondences on the trajectory attitude (Sec. 5.1.1) and the point-cloud geo-referencing error (Sec. 5.1.2) and (ii) the extraction of 2D-3D correspondences from the existing (previously trained) networks.

### 5.1 Emulated correspondences

For this work, the locations of the extracted points were randomly selected to be uniformly distributed in the study area. We selected $\approx 50$ pairs of 2D-3D correspondences per image, along with $\approx 100 - 200$ image tie-points (2D-2D) per image and $\approx 4$ lidar correspondences (3D-3D) every 30 m assuming a uniform spatial distribution. The 2D-3D correspondences were considered with sub-pixel prior uncertainty.

#### 5.1.1 Impact on trajectory attitude

When incorporating the emulated pixel-to-point correspondences (Sec. 4.2) into DN with low-cost IMU measurements (case C), the impact on the quality of the trajectory attitude is significant. Similar improvement is observed when using these correspondences together with image-to-image and lidar-to-lidar correspondences (case D). These results are compared to the previously reported cases A and B. This improvement is illustrated in Fig. 4 with the corresponding statistics provided in Tab. 2, showing the attitude error distribution per attitude component [roll, pitch, yaw] for all study cases. The attitude error is computed with respect to the reference trajectory and reflects the deviation of each trajectory solution from it.

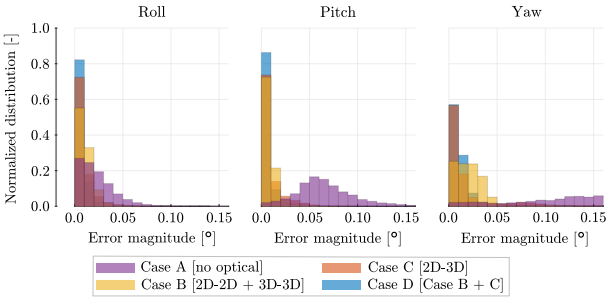


Figure 4. Normalized distribution of the ALS trajectory attitude error in degrees. The error of each trajectory corresponds to its deviation from the high-quality trajectory.

It is observed that the use of 2D-3D constraints (cases C and D) controls the attitude solution more significantly than the previously reported case (B), reducing its error and dispersion in all three components. In other words, the solution with single-domain correspondences (case B) is rather weak in the yaw component, while the use of cross-domain correspondences limits its drift, indicating the merit of this new type of constraints.

#### 5.1.2 Impact on point-cloud geo-referencing

Given the DN computed trajectories and the formerly calibrated lidar mounting parameters, we geo-reference the laser vectors and compare them with the reference point-cloud (Sec. 4.4) to study the effect of the new type of spatial constraints in DN. The trajectory improvement observed in Sec. 5.1.1 when using pixel-to-point correspondences (C) is also reflected in the reduced geo-referencing error, in all studied flight lines as depicted in Fig. 5.

| Traj. # | Case A | Case B | Case C | Case D |
|---|---|---|---|---|
| R [°] | | | | |
| MEAN | 0.003 | **-0.001** | -0.003 | **-0.001** |
| STD | 0.033 | 0.014 | 0.013 | **0.010** |
| RMSE | 0.033 | 0.014 | 0.013 | **0.010** |
| P [°] | | | | |
| MEAN | -0.067 | -0.001 | -0.002 | **0.000** |
| STD | 0.035 | 0.012 | 0.015 | **0.010** |
| RMSE | 0.075 | 0.012 | 0.015 | **0.010** |
| Y [°] | | | | |
| MEAN | 0.066 | 0.013 | -0.012 | **0.003** |
| STD | 0.169 | 0.029 | 0.056 | **0.025** |
| RMSE | 0.182 | 0.032 | 0.057 | **0.025** |

Table 2. Error statistics of the estimated trajectories. The best trajectory for each metric (line) is highlighted in bold.

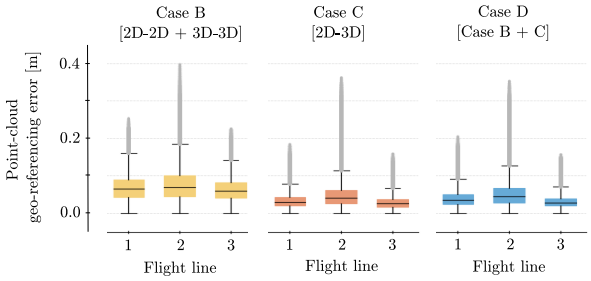


Figure 5. Geo-referencing error (norm) of the lidar point-clouds per flight line, with and without the pixel-to-point (2D-3D) constraints.

Indeed, as shown in the graphs of cases C and D where the 2D-3D constraints are involved, the mean geo-referencing error is reduced by $30 - 50\%$ compared to case B where only single-domain optical constraints are used. Additionally, there is a small reduction in the maximum error and the error dispersion indicating the increased mapping accuracy. The difference between the last two graphs (C and D) is however minimal.

### 5.2 Correspondences via existing networks

We consider the following metrics to evaluate the performance of the pre-trained SOTA networks on the studied aerial dataset, which are summarized in Tab. 3.

- N° of correct matches: average number of correct pixel-to-point matches per image, where the distance between a 2D pixel and the re-projection of its paired 3D point is below a threshold given its ground-truth projection. To facilitate comparison in this section, we express the pixel error threshold in meters in the 3D space given the down-sampled GSD (Sec. 4.3), i.e. 2 $m$ for both detection-free networks.
- Inlier ratio: average proportion of correct matches relative to the total number of output matches per image.

**Specificity:** The threshold mentioned in the first metric of n° of correct matches is used only for D-GLSNet and VP2P-Match, where the input datasets are whole images and point-clouds, as

| Metric - (average per image) | LCD | D-GLSNet | VP2P-Match |
|---|---|---|---|
| N° correct matches | **18.6** | 1.3 | 0.8 |
| Inlier ratio | **1.6**% | 0.78% | 0.70% |

Table 3. Performance of the existing networks on the aerial dataset. For each metric, we have the average of all images.

opposed to pixel and point patches. Thus, all pixels and points present in the datasets are potential candidates. In LCD, the input dataset is small non-overlapping patches centered in specific pixels and their corresponding 3D points. Thus, correct or wrong matches can be identified based on their indices, while their re-projection errors are multipliers of the constant patch size used in data creation.

**Discussion:** Despite not being end-to-end (i.e. the pixel-to-point correspondence extraction is the main task of the network and not an intermediate step), the detect-then-match network LCD provided the best results outperforming the other networks, that struggled to extract meaningful 2D-3D correspondences. The average number of correct matches per image with LCD indicates that the network can extract some meaningful information. However, all three networks failed to generalize effectively on new data, given their low inlier ratio scores. More significant information could be potentially retrieved with some further refinement, e.g. with PnP-RANSAC (Perspective-n-Point RANdom SAmple Consensus) filtering (Fischler and Bolles, 1981, Wu and Hu, 2006).

**Limitations:** The size and resolution of the studied scenes in both images and lidar data are challenging for detection-free networks due to the limited input size and number of points they can handle. Detection-free networks, while being more performant on specific datasets as indicated in the literature, are more complex and thus harder to retrain and fine-tune. This complexity often results in difficulties generalizing to new data, which was evident in our evaluations as airborne data were not used for training.

## 6. Conclusions & outlook

In this work, we have proposed to incorporate a new cross-domain observation model that leverages matches between images and lidar point-clouds (so-called pixel-to-point correspondences) and fuses them along with GNSS and raw inertial measurements in the DN adjustment for optimal estimation of trajectory and other parameters. We perform the proof of concept by employing emulated correspondences from a controlled aerial dataset to showcase their potential benefit in the improvement of the trajectory quality and subsequently the mapping quality. We also evaluated the performance of existing open-source deep-learning architectures in extracting real pixel-to-point correspondences from the same aerial dataset.

Using emulated correspondences between images and lidar point-clouds in the DN provided encouraging indications for attitude improvement, especially in the yaw component, in comparison to the previously described baseline of single-domain correspondences (Mouzakidou et al., 2024). In turn, this improved the geo-referencing accuracy of the lidar point-cloud by $30 - 50\%$. Further studies will focus on the amelioration of other parameters related to sensor and system calibration, e.g. parameter observability as boresight angles. On the other hand, we show that the current deep-learning architectures, that were trained on other than airborne data, cannot extract direct pixel-to-point correspondences at sufficient number and quality.

This evidence gives research prospects for advancing the cross-domain (image-to-lidar) matching workflow. This includes the need to retrain the networks on aerial datasets, possibly even evolving the architecture in terms of the input details, to obtain reliable cross-modality descriptors. If realized practically, these constraints may be fundamental in establishing a common and qualitatively better data-alignment workflow for hybrid sensors in aerial mapping.

## 7. Acknowledgments

The helicopter flight and the reference sensors were provided by Helimap Systems. The personal involvement and contribution of its director Dr. Julien Vallet are highly appreciated. This contribution is supported by the Swiss Innovation Agency projects: Innosuisse (i) no. 53622.1 and (ii) no. 119.293.

## References

Brown, D., 1971. Close-Range Camera Calibration. *Photogrammetric Engineering*, 37(8), 855–866.

Brun, A., Cucci, D. A., Skaloud, J., 2022. Lidar point–to–point correspondences for rigorous registration of kinematic scanning in dynamic networks. *ISPRS Journal of Photogr. and Rem. Sens.*, 189, 185–200. doi:10.1016/j.isprsjprs.2022.04.027.

Charles, R. Q., Su, H., Kaichun, M., Guibas, L. J., 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 77–85. doi:10.1109/CVPR.2017.16.

CityMapper - Leica Geosystems, 2024. Leica CityMapper-2: Hybrid Sensors for Airborne Systems. Accessed: 2024-12-02.

Colomina, I., Gimenez, M., Rosales, J., Wis, M., Gomez, A., Miguelsanz, P., 2004. Redundant IMUs for Precise trajectory determination. *ISPRS Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 34, Part B(Commision 1), 7.

Cucci, D. A., Rehak, M., Skaloud, J., 2017. Bundle adjustment with raw inertial observations in UAV applications. *ISPRS Journal of Photogr. and Rem. Sens.*, 130, 1-12. doi:10.1016/j.isprsjprs.2017.05.008.

Cucci, D. A., Skaloud, J., 2019. On Raw Inertial Measurements In Dynamic Networks. *ISPRS Annals of the Photogr., Rem. Sens. and Spat. Inform. Sci.*, IV-2/W5, 549–557. doi:10.5194/isprs-annals-IV-2-W5-549-2019.

Feng, M., Hu, S., Ang, M. H., Lee, G. H., 2019. 2D3D-Matchnet: Learning To Match Keypoints Across 2D Image And 3D Point Cloud. *2019 International Conference on Robotics and Automation (ICRA)*, 4790–4796. doi:10.1109/ICRA.2019.8794415.

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6), 381–395. doi:10.1145/358669.358692.

Glennie, C. L., 2007. Rigorous 3D error analysis of kinematic scanning LIDAR systems. *Journal of Applied Geodesy*, 1, 147-157. doi:10.1515/jag.2007.017.

Glira, P., Pfeifer, N., Mandlburger, G., 2019. Hybrid orientation of airborne LiDAR point clouds and aerial images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W5, 567–574. doi:10.5194/isprs-annals-IV-2-W5-567-2019.

Hussnain, Z., Elberink, S. O., Vosselman, G., 2021. Enhanced trajectory estimation of mobile laser scanners using aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173, 66-78. doi:10.1016/j.isprsjprs.2021.01.005.

Lazebnik, S., Schmid, C., Ponce, J., 2005. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1265–1278. doi:10.1109/TPAMI.2005.151.

Li, L., Ma, Y., Tang, K., Zhao, X., Chen, C., Huang, J., Mei, J., Liu, Y., 2023a. Geo-Localization With Transformer-Based 2D-3D Match Network. *IEEE Robotics and Automation Letters*, 8(8), 4855–4862. doi:10.1109/LRA.2023.3290526.

Li, M., Qin, Z., Gao, Z., Yi, R., Zhu, C., Guo, Y., Xu, K., 2023b. 2D3D-MATR: 2D-3D Matching Transformer for Detection-free Registration between Images and Point Clouds. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 14082–14092. doi:10.1109/ICCV51070.2023.01299.

Liu, W., Lai, B., Wang, C., Bian, X., Yang, W., Xia, Y., Lin, X., Lai, S.-H., Weng, D., Li, J., 2020. Learning to Match 2D Images and 3D LiDAR Point Clouds for Outdoor Augmented Reality. *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 654–655. doi:10.1109/VRW50115.2020.00178.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

Mouzakidou, K., Brun, A., Cucci, D. A., Skaloud, J., 2024. Airborne sensor fusion: Expected accuracy and behavior of a concurrent adjustment. *ISPRS Open Journal of Photogr. and Rem. Sens.*, 12, 100057. doi:10.1016/j.ophoto.2023.100057.

Mouzakidou, K., Cucci, D. A., Skaloud, J., 2022. On the benefit of concurrent adjustment of active and passive optical sensors with GNSS & raw inertial data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-1-2022, 161–168. doi:10.5194/isprs-annals-V-1-2022-161-2022.

Nadeem, U., Bennamoun, M., Togneri, R., Sohel, F., Miri Rekavandi, A., Boussaid, F., 2023. Cross domain 2D-3D descriptor matching for unconstrained 6-DOF pose estimation. *Pattern Recognition*, 142, 109655. doi:10.1016/j.patcog.2023.109655.

Pham, Q.-H., Uy, M. A., Hua, B.-S., Nguyen, D. T., Roig, G., Yeung, S.-K., 2020. LCD: Learned Cross-Domain Descriptors for 2D-3D Matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 11856–11864. doi:10.1609/aaai.v34i07.6859.

Pöppl, F., Neuner, H., Mandlburger, G., Pfeifer, N., 2023. Integrated trajectory estimation for 3D kinematic mapping with GNSS, INS and imaging sensors: A framework and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196, 287–305. doi:10.1016/j.isprsjprs.2022.12.022.

Pöppl, F., Ullrich, A., Mandlburger, G., Pfeifer, N., 2024. A flexible trajectory estimation methodology for kinematic laser scanning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215, 62-79. doi:10.1016/j.isprsjprs.2024.06.014.

Ren, S., Zeng, Y., Hou, J., Chen, X., 2023. CorrI2P: Deep Image-to-Point Cloud Registration via Dense Correspondence. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3), 1198–1208. doi:10.1109/TCSVT.2022.3208859.

Rusu, R. B., Cousins, S., 2011. 3D is here: Point Cloud Library (PCL). *2011 IEEE International Conference on Robotics and Automation*, 1–4. doi:10.1109/ICRA.2011.5980567.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning Feature Matching With Graph Neural Networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Seattle, WA, USA, 4937–4946. doi:10.1109/CVPR42600.2020.00499.

Vallet, J., Gressin, A., Clausen, P., Skaloud, J., 2020. Airborne and Mobile LiDAR, which Sensors for which Application? *ISPRS Archives of the Photogr., Rem. Sens. and Spat. Inform. Sci.*, 43B1, 397-405. doi:10.5194/isprs-archives-XLIII-B1-2020-397-2020.

Wang, B., Chen, C., Cui, Z., Qin, J., Lu, C. X., Yu, Z., Zhao, P., Dong, Z., Zhu, F., Trigoni, N., Markham, A., 2021. P2-Net: Joint Description and Detection of Local Features for Pixel and Point Matching. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 15984–15993. doi:10.1109/ICCV48922.2021.01570.

Wu, Y., Hu, Z., 2006. PnP Problem Revisited. *Journal of Mathematical Imaging and Vision*, 24, 131–141. doi:10.1007/s10851-005-3617-z.

Yao, G., Xuan, Y., Chen, Y., Pan, Y., 2024. Quantity-Aware Coarse-to-Fine Correspondence for Image-to-Point Cloud Registration. arXiv:2307.07142 [cs].

Zhong, Y., 2009. Intrinsic shape signatures: A shape descriptor for 3D object recognition. *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, 689-696. doi:10.1109/ICCVW.2009.5457637.

Zhou, J., Ma, B., Zhang, W., Fang, Y., Liu, Y.-S., Han, Z., 2023. Differentiable Registration of Images and LiDAR Point Clouds with VoxelPoint-to-Pixel Matching. *arXiv*. doi:10.48550/arXiv.2312.04060.