Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing

Armand Zampieri, Guillaume Charpiat, Nicolas Girard, Yuliya Tarabalka

Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing

Armand Zampieri[1], Guillaume Charpiat[2], Nicolas Girard[1], and Yuliya Tarabalka[1]

[1] TITANE team, INRIA, Université Côte d'Azur, France
[2] TAU team, INRIA, LRI, Université Paris-Sud, France
firstname.lastname@inria.fr

Abstract. We tackle here the problem of multimodal image non-rigid registration, which is of prime importance in remote sensing and medical imaging. The difficulties encountered by classical registration approaches include feature design and slow optimization by gradient descent. By analyzing these methods, we note the significance of the notion of scale. We design easy-to-train, fully-convolutional neural networks able to learn scale-specific features. Once chained appropriately, they perform global registration in linear time, getting rid of gradient descent schemes by predicting directly the deformation.
We show their performance in terms of quality and speed through various tasks of remote sensing multimodal image alignment. In particular, we are able to register correctly cadastral maps of buildings as well as road polylines onto RGB images, and outperform current keypoint matching methods.

Keywords: Multimodal · Alignment · Registration · Remote sensing

1 Introduction

Image alignment, also named *non-rigid registration*, is the task of finding a correspondence field between two given images, *i.e.* a deformation which, when applied to the first image, warps it to the second one. Such warpings can prove useful in many situations: to transfer information between several images (for instance, from a template image with labeled parts), to compare the appearance of similar parts (as pixel intensity comparison makes sense only after alignment), or to estimate spatial changes (to monitor the evolution of a tumor given a sequence of scans of the same patient over time, for instance). Image alignment has thus been a predominant topic in fields such as medical imaging or remote sensing [33,20].

1.1 Remote sensing & Image alignment

In remote sensing, images of the Earth can be acquired through different types of sensors, in the visible spectrum or not, from satellites or planes, with various

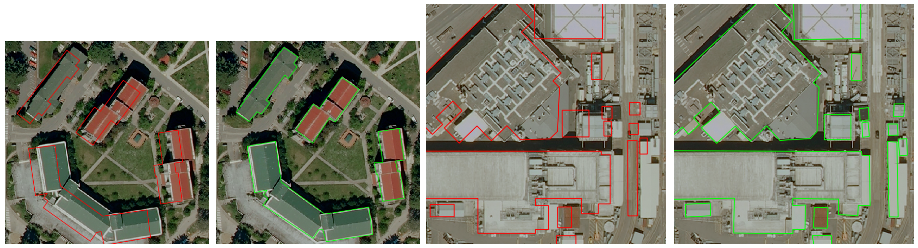


Fig. 1. Examples of multimodal image alignment. We align aerial images with cadastral images. For each example: Left: original image with OpenStreetMap (OSM) map (in red), right: after our realignment (result in green).

spatial precision (from cm to km range). The analysis of these images allows the monitoring of ecosystems (plants [11], animals [35]...) and their evolution (drought monitoring, natural disasters and associated help planning), urban growth, as well as the automatic creation of maps [21] or more generally digitizing the Earth.

However, the geographic localization of pixels in these images is limited by a number of factors, such as the positioning precision and the effect of the relief on non-vertical points of view. The deformation of these images is significant: for instance, in OpenStreetMap [10], objects may be shifted by 8 meters (which is far above the required precision of maps for autonomous driving), which means an error displacement of more than 20 pixels for a 30 cm/pixel resolution.

These deformations prevent a proper exploitation of such data. For instance, let us consider the task of finding buildings and roads in a remote sensing image. While ground truth is actually available in considerable amounts, such as in OpenStreetMap (OSM) based on cadastral information, which gives coordinates (latitude and longitude) of each building corner, this hand-made ground truth is often inaccurate, because of human mistakes. Thus it is not possible to learn from it, as remote sensing images are not properly aligned to it and objects might even not overlap. This is a severe issue for the remote sensing field in the era of big data and machine learning. Many works have been focusing on this problem [2], from the use of relief knowledge to dedicated hand-designed alignment algorithms. Another approach worth mentioning is to train coarsely on the datasets available and fine-tune on small better-hand-aligned datasets [18]. We will here tackle the problem of non-rigid alignment directly.

1.2 Classical approaches for non-rigid registration

Tasks. Image registration deals with images either of the same modality (same sensor), or not. When of the same modality, the task is typically to align different but similar objects (*e.g.*, faces [4] or organs of different people [13]), or to align the same object but taken at different times (as in the tumor monitoring example). On the other side, multi-modal registration deals with aligning images usually of the same object but seen by different sensors, which capture different physical properties, at possibly different resolutions. For instance in medical imaging, MR and CT scans capture the density of water and of matter respectively, while in remote sensing RGB and hyperspectral data capture information from different light frequencies (infrared, etc.). In our case of study, we focus on the alignment of RGB remote sensing images with cadastres, *i.e.* vector-format images with polygonal representations of all buildings and roads, hand-made by local authorities, map makers or OpenStreetMap users as in Figures 1 and 2.

Whether mono-modal or multi-modal, image registration faces two challenges: first, to describe locally image data, and then, to match points with similar description, in a spatially coherent manner. Historically, two main classical approaches have emerged:

Matching key-points. The first one consists in sampling a few key-points from each image (*e.g.* with Harris corner detection), in describing them locally (with SIFT, SURF, HOG descriptors...) [32,7], in matching these points [27] and then interpolating to the rest of the image. The question is then how to design proper sampling criteria, descriptors, and matching algorithm. In the multi-modal case, one would also have to design or learn the correspondence between the descriptors of the two modalities. Note that high-precision registration requires a dense sampling, as well as consequently finer descriptors.

Estimating a deformation field by gradient descent. The second approach, particularly popular in medical imaging, consists in estimating a dense deformation field from one image to the other one [1,9,26,15,13]. One of its advantages over the first approach is to be able to model objects, to make use of shape statistics, etc. The warping is modeled as a smooth vector field ϕ, mapping one image domain onto the other one. Given two images I_1 and I_2, a criterion $C(I_1 \circ \phi, I_2)$ is defined, to express the similarity between the warped image $I_1 \circ \phi$ and the target I_2, and is optimized with respect to ϕ by gradient descent. Selecting a suitable similarity criterion C is crucial, as well as designing carefully the gradient descent, as we will detail in section 2.

1.3 The new paradigm: neural networks

The difficulty to design or pick particular local descriptors or matching criteria among many possibilities is the trait of computer vision problems where the introduction of neural networks can prove useful. The question is how. Machine learning techniques have already been explored to learn similarity measures between different imaging modalities [42], for instance using kernel methods to register MR and CT brain scans [17], or very recently with neural networks [30,16,39], but without tackling the question of scale. We will aim at designing a system able to learn scale-specific and modality-specific features, and able to perform multimodal image registration densely and swiftly, without the use of any iterative process such as gradient descent which hampers classical approaches. Our contributions are thus:

- a swift system to register images densely,

– learning features to register images of different modalities,
– learning scale-specific features and managing scales,
– designing a (relatively small) neural network to do this end-to-end,
– aligning remote sensing images with cadastral maps (buildings and roads),
– providing a long-awaited tool to create large-scale benchmarks in remote sensing.

We first analyze the problems related to scale when aligning images, in order to design a suitable neural network architecture. We show results on benchmarks and present additional experiments to show the flexibility of the approach.

2 Analysis of the gradient descent framework

In order to analyze issues that arise when aligning images, let us first consider the case of mono-modal registration, for simplicity. Keeping the notations from Section 1.2, we pursue the quest for a reasonable criterion $C(I_1 \circ \phi, I_2)$ to optimize by gradient descent to estimate the deformation ϕ.

2.1 A basic example

Too local quantities such as the pixellic intensity difference $C(I_1 \circ \phi, I_2) = \|I_1 \circ \phi - I_2\|^2_{L^2}$ would create many local minima and get the gradient descent stuck very fast. Indeed, if as a toy example one considers I_1 and I_2 to be two white images with a unique black dot at different locations $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively, the derivative of $C(I_1 \circ \phi, I_2)$ with respect to ϕ will never involve quantities based on these two points at the same time, which prevents them from being influenced by each other:

$$\frac{\partial C(I_1 \circ \phi, I_2)}{\partial \phi}(\mathbf{x}) = 2\big(I_1 \circ \phi(\mathbf{x}) - I_2(\mathbf{x})\big)\ (\nabla_{\mathbf{x}} I_1)_{(\phi(\mathbf{x}))}$$

is 0 at all points $\mathbf{x} \neq \mathbf{x}_1$, and at $\mathbf{x}_1$ the deformation ϕ (initialized to the identity) evolves to make it disappear from the cost C by shrinking the image around. Thus the derivative of the similarity cost C with respect to the deformation ϕ does not convey any information pushing $\mathbf{x}_1$ towards $\mathbf{x}_2$, but on the contrary will make the descent gradient stuck in this (very poor) local minimum.

Instead of the intensity $I(\mathbf{x})$, one might want to consider other local, higher-level features $L(I)(\mathbf{x})$ such as edge detectors, in order to grasp more meaningful information, and thus minimize a criterion for instance of the form:

$$C(I_1 \circ \phi, I_2) = \|L(I_1 \circ \phi) - L(I_2)\|^2_{L^2} \,. \tag{1}$$

2.2 Neighborhood size

The solution consists in considering local descriptors involving larger spatial neighborhoods, wide enough so that the image domains involved in the computations of $L(I_1 \circ \phi)(\mathbf{x}_1)$ and $L(I_2)(\mathbf{x}_2)$ for two points $\mathbf{x}_1$ and $\mathbf{x}_2$ to be matched overlap significantly. For instance, the computation of the Canny edge detector is performed over a truncated Gaussian neighborhood, whose size is pre-defined by the standard deviation parameter σ. Another example is the local-cross correlation, which compares the local variations of the intensity around $\mathbf{x}_1$ and $\mathbf{x}_2$ within a neighborhood of pre-defined size [13]. Another famous example is the mutual information between the histograms of the intensity within a certain window with pre-defined size.

2.3 Adapting the scale

In all these cases, the neighborhood size is particularly important: if too small, the gradient descent will get stuck in a poor local minimum, while if too large, the image details might be lost, preventing fine registration. What is actually needed is this neighborhood size to be of the same order of magnitude as the displacement to be found. As this displacement is unknown, the neighborhood size needs to be wide enough during the first gradient steps (possibly covering the full image), and has to decrease with time, for the registration to be able to get finer and finally reach pixellic precision. Controlling the speed of this decrease is a difficult matter, leading to slow optimization. Moreover, the performance of the descriptors may depend on the scale, and different descriptors might need to be chosen for the coarse initial registration than for the finest final one. In addition to the difficult task of designing [43,40] or learning [17] relevant descriptors L_s for each scale, this raises another issue, that the criterion C_s to optimize

$$C_s(I_1 \circ \phi,\, I_2) = \|L_s(I_1 \circ \phi) - L_s(I_2)\|^2_{L^2} \tag{2}$$

now depends on the current neighborhood size $s(t)$, which is itself time-dependent, and thus the optimized criterion $C_{s(t)}$ might increase when the descriptor $L_{s(t)}$ evolves: the optimization process is then not a gradient descent anymore.

One might think of scale-invariant descriptors such as SIFT, however the issue is not just to adapt the scale to a particular location within an image, but to adapt it to the amplitude of the deformation that remains to be done to be matched to the *other* image.

2.4 Multi-resolution viewpoint

Another point of view on this scale-increasing process is to consider that the descriptors and optimization process remain the same at all scales, but that the *resolution* of the image is increasing. The algorithm is then a loop over successive resolutions [13,4], starting from a low-resolution version of the image, waiting for convergence of the gradient descent, then upsampling the deformation field found to a higher resolution version of the image, and iterating until the original resolution is reached. The limitation is then that the same descriptor has to be used for each resolution, and, as previously, that the convergence of a gradient descent has to be reached at each scale, leading to slow optimization. A different approach consists in dealing with all scales simultaneously by considering a multi-scale parameterisation of the deformation [31]. However, the same local minimum problem would be encountered if implemented naively; heuristics then need to be used to estimate at which scale the optimization has currently to be performed locally.

2.5 Keeping deformations smooth

Deformations are usually modeled as diffeomorphisms [1,9,15,13], *i.e.* smooth one-to-one vector fields, in order to avoid deleting image parts. The smoothness is controlled by an additional criterion to optimize, quantifying the regularity of the deformation ϕ, such as its Sobolev norm (penalizing fast variations). As in any machine learning technique, this regularity term sets a prior over the space of possible functions (here, deformations), preventing overfitting (here, spatial noise). But once again, the smoothness level required should depend on the scale, *e.g.* prioritizing global translations and rotations at first, while allowing very local moves when converging. This can be handled by suitable metrics on instantaneous deformations [5,34]; yet in practice these metrics tend to slow down convergence by over-smoothing gradients $\nabla_\phi C$ at finest scales.

3 Introducing neural networks

3.1 Learning iterative processes

As neural networks have proved useful to replace hand-designed features for various tasks in the literature recently, and convolutional ones (CNN) in particular in computer vision, one could think, for mono-modal image alignment, of training a CNN in the Siamese network setup [3,6], in order to learn a relevant distance between image patches. The multi-modal version of this would consist in training two CNN (one per modality) with same output size, in computing the Euclidean norm of the difference of their outputs as a dissimilarity measure, and in using that quantity within a standard non-rigid alignment algorithm, such as a gradient descent over (1). For training, this would however require to be able to differentiate the result of this iterative alignment process with respect to the features. This is not realistic, given the varying, usually large number of steps required for typical alignment tasks. A similar approach was nonetheless successfully used in [18], for the simpler task of correcting blurry segmentation maps, sharpening them and relying on image edges. For this, a partial differential equation (PDE) was mimicked with a recurrent network, and the number of steps applying this PDE was pre-defined to a small value (5), sufficient for that particular problem. In the same spirit, for image denoising, in [22,37] the proximal operator used during an iterative optimization process is modeled by a neural network and learned. In [25], the Siamese network idea is used, but for matching only very few points. It is also worth noting that, much earlier, in [17], a similarity criterion between different modalities was learned, with a kernel method, but for rigid registration only.

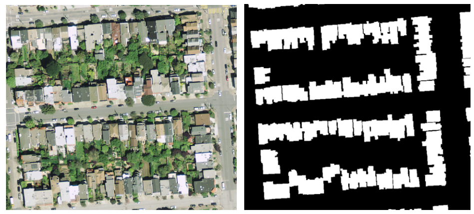


Fig. 2. Multimodal pair of satisfyingly aligned images, from the database. Left: aerial RGB image, right: vector-format cadastral image (buildings are shown in white).

3.2 A more direct approach

As seen in the previous sections, aligning images thanks to a gradient descent over the deformation ϕ has the following drawbacks: it is slow because of the need to ensure convergence at each scale, it is actually not a real gradient descent if descriptors are scale-dependent, and it induces a long backpropagation chain when learning the descriptors. To get rid of this iterative process, we propose to predict directly its final result at convergence. That is, given images I_1 and I_2, to predict directly the optimal deformation ϕ so that $I_1 \circ \phi$ and I_2 are aligned. Also, instead of proceeding in two steps: first learning the features L required to define the criterion C in (1), then finding the deformation ϕ minimizing C, we propose to directly learn the deformation as in a standard machine learning setup, that is, from examples. Given a training set of input pairs $P = (I_1, I_2)$ together with the expected associated output ϕ_P, we aim at learning the function $P \mapsto \phi_P$.

3.3 Machine learning setting

Training set. We first consider the task of aligning geolocalized aerial RGB images with binary maps from OpenStreetMap indicating building locations. As explained in Section 1.1, the matching is usually imperfect. Creating the deformation ground truth by manually performing the warpings would be too time-consuming. Instead, we extract image pairs which visually look already well aligned, as in Figure 2. This way we obtain a dataset composed of 5000×5000 image pairs (aerial RGB image, binary vector-format building map) at resolution 0.3m/pixel, for which the deformation ϕ to be found is the identity.

We generate an artificial training set by applying random deformations to the cadastral vectorial maps, moving accordingly the corners of the polygons it contains, and then generating the new binary maps by rasterization. We thus obtain a training set of pairs of non-registered images, with known deformations. As typical deformations in reality are smooth, we model our family of random deformations as: a global translation $\mathbf{v}_0$ taken uniformly within a certain range $[-r, +r]^2$, plus a mixture of Gaussian functions with random shifts $\mathbf{v}_i$, centers $\mathbf{x}_i$ and covariance matrices S_i:

$$\phi(\mathbf{x}) \;=\; \mathbf{v}_0 \,+\, \sum_{i=1}^{n} \mathbf{v}_i \; e^{-(\mathbf{x}-\mathbf{x}_i)\, S_i \,(\mathbf{x}-\mathbf{x}_i)} \tag{3}$$

with uniformly random $\mathbf{v}_i$, S_i, $\mathbf{x}_i$ within suitable pre-defined ranges (S_i being symmetric positive definite). This way, we can drastically augment the dataset by applying arbitrarily many random deformations to initially well-aligned images.

Optimization criterion. The loss considered is simply the Euclidean norm of the prediction error:

$$C(\mathbf{w}) = \underset{(I_1, I_2, \phi_{\mathrm{GT}}) \in \mathcal{D}}{\mathbb{E}} \left[\sum_{\mathbf{x} \in \Omega(I_2)} \left\| \widehat{\phi}_{(\mathbf{w})(I_1,I_2)}(\mathbf{x}) - \phi_{\mathrm{GT}}(\mathbf{x}) \right\|_2^2 \right]$$

i.e. the expectation, over the ground truth dataset $\mathcal{D}$ of triplet examples (RGB image I_1, cadastral image I_2, associated deformation ϕ_{GT}), of the sum, over all pixels $\mathbf{x}$ in the image domain $\Omega(I_2)$, of the norm of the difference between the ground truth deformation $\phi_{\mathrm{GT}}(\mathbf{x})$ and the one predicted $\widehat{\phi}_{(\mathbf{w})(I_1,I_2)}(\mathbf{x})$ for the pair of images (I_1, I_2) given model parameters $\mathbf{w}$ (*i.e.* the neural network weights). In order to make sure that predictions are smooth, we also consider for each pixel a penalty over the norm of the (spatial) Laplacian of the deformation:

$$\left\| \triangle \widehat{\phi}_{(\mathbf{w})(I_1,I_2)}(\mathbf{x}) \right\|_2^2 \tag{4}$$

which penalizes all but affine warpings. In practice in the discrete setting this sum is the deviation of $\widehat{\phi}(\mathbf{x})$ from the average over the 4 neighboring pixels: $\left\| \widehat{\phi}(\mathbf{x}) - \frac{1}{4} \sum_{\mathbf{x}' \sim \mathbf{x}} \widehat{\phi}(\mathbf{x}') \right\|_2^2$.

3.4 A first try

We first produce a training set typical of real deformations by picking a realistic range $r = \pm 20$ pixels of deformation amplitudes. We consider a fully-convolutional neural network, consisting of two convolutional networks (one for each input image I_i), whose final outputs are concatenated and sent to more convolutional layers. The last layer has two features, *i.e.* emits two real values per pixel, which are interpreted as $\widehat{\phi}(\mathbf{x})$. In our experiments, such a network does not succeed in learning deformations: it constantly outputs $\widehat{\phi}(\mathbf{x}) = (0,0)\ \forall \mathbf{x}$, which is the best constant value for our loss, *i.e.* the best answer one can make when not understanding the link between the input (I_1, I_2) and the output ϕ for a quadratic loss: the average expected answer $\mathbb{E}_{(I_1,I_2,\phi_{\mathrm{GT}})\in\mathcal{D}}[\phi]$, which is $(0,0)$ in our case.

We also tried changing representation by predicting bin probabilities $p\big(\Phi_x(\mathbf{x}) \in [a, a+1]\big)$, $p\big(\Phi_y(\mathbf{x}) \in [b, b+1]\big)$ for each integer $-r \leqslant a, b < r$, by outputting 2 vectors of $2r$ real values per pixel, but this lead to the same result.

3.5 Dealing with a single scale

The task in Section 3.4 is indeed too hard: the network needs to develop local descriptors at all scales to capture all information, and is asked to perform a fine matching with $(2r)^2 \simeq 1700$ possibilities for each pixel $\mathbf{x}$.

This task can be drastically simplified though, by requiring the network to perform the alignment *at one scale s only*. By this, we mean:

<u>Task at scale s:</u> *Solve the alignment problem for the image pair* (I_1, I_2), *with a precision required of* $\pm 2^s$ *pixels, under the assumption that the amplitude of the registration to be found is not larger than* 2^{s+1} *pixels.*

For instance, at scale $s = 0$, the task is to search for a pixelwise precise registration (± 1 pixel) on a dataset prepared as previously but with amplitude $r = 2^{s+1} = 2$. As a first approximate test, we train the same network as described earlier, with $r = 2$, *i.e.* each of the 2 coordinates of $\phi(\mathbf{x})$ take value in $[-2, 2]$, and we consider a prediction $\widehat{\phi}(\mathbf{x})$ to be correct if in the same unit-sized bin as $\phi(\mathbf{x})$. Without tuning the architecture or the optimization method, we obtain, from the first training run, about 90% of accuracy, to be compared to $\sim 6\%$ for a random guess.

Thus, it is feasible, and easy, to extract information when specifying the scale. Intuitively, the search space is much smaller; in the light of Section 2.2, the descriptor receptive field required for such a ± 1 pixel task is just of radius 1. And indeed, in the classical framework for mono-modal registration, a feature as simple as the image intensity would define a sufficiently good criterion (1), as the associated gradient step involves the comparison to the next pixel (through $\nabla_{\mathbf{x}} I_1$). Note that such a simple intensity-based criterion would not be expected to perform more, *e.g.* find deformations of amplitude $r \geqslant 2$ pixels in the general case (textures).

Designing a suitable neural network architecture. We now propose better architectures to solve that alignment task at scale $s = 0$. We need a fully-convolutional architecture since the output is a 2-channel image of the same size as the input, and we need to go across several scales in order to understand

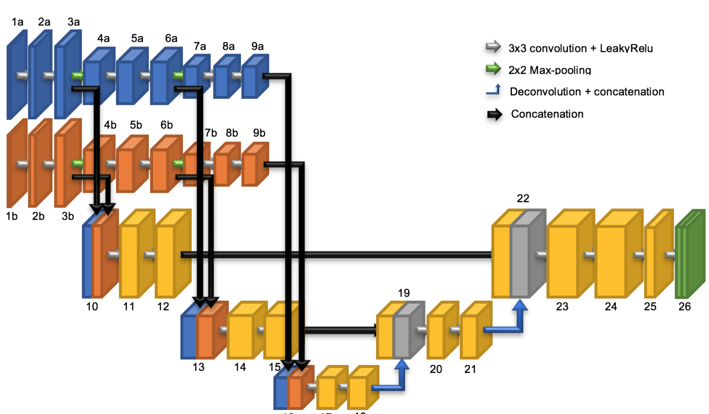


Fig. 3. Network architecture for one scale. The two input images I_1 and I_2 are fed to layers 1a and 1b respectively. The output is a 2 dimensional vector map (layer 26 with 2 channels). See supplementary materials for all details.

which kind of object part each pixel belongs to, in each modality. High-level features require a wide receptive field, and are usually obtained after several pooling layers, in a pyramidally shaped network. The output needs however to be of the same resolution as the input, which leads to autoencoder-like shapes. In order to keep all low-level information until the end, and not to lose precision, we link same-resolution layers together, thus obtaining a kind of double-U-net network (U-nets [36] were developed for medical image segmentation). As the 2 input images are not registered, and to get modality-specific features, we build 2 separate convolutional pyramids, one for each modality (in a similar fashion as networks for stereo matching [44]), but concatenate their activities once per scale to feed the double U-net. The architecture is summarized in Figure 3. The network is trained successfully to solve the $s = 0$ task as explained previously.

3.6 A chain of scale-specific neural networks

We now solve the general alignment task very simply:

Solution for task at scale s: *Downsample the images by a factor 2^s; solve the alignment task at scale 0 for these reduced images, and upsample the result with the same factor.*

Full alignment algorithm: *Given an image pair (I_1, I_2) of width w, iteratively solve the alignment task at scale s, from $s = \log_2 w$ until $s = 0$.*

One can choose to use the same network for all scales, or different ones if we expect specific features at each scale, as in remote sensing or medical imaging.

The full processing chain is shown in Figure 4. Note a certain global similarity with ResNet [12], in that we have a chain of consecutive scale-specific blocks, each of which refining the previously-estimated deformation, not by adding to it but by diffeomorphism composition: $\phi_{s-1} = \phi_s \circ \big(\mathrm{Id} + f(I_1 \circ \phi_s, I_2 \circ \phi_s)\big)$. Another difference with ResNet is that we train each scale-specific block independently, which is much easier than training the whole chain at once. A similar idea was also independently developed in [28] for optical flow estimation; their architecture is more complex in that the input to each block is not only the downsampled images but also the flow from the previous resolution, and the flows are added instead of composed. This lead to much higher training times (days instead of hours for us). Also, obviously, [28] cannot deal with multimodality.

Note that the overall complexity of an alignment is very low, linear in the image size. Indeed, for a given image with n pixels, a similar convolutional architecture is applied to all reduced versions by factors 2^s, of size $2^{-s} \times 2^{-s} n$ pixels, leading to a total cost of $n(1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots)K < \frac{4}{3} nK$ where K is the constant per-pixel convolutional cost. This is to be compared with the classical gradient descent based approaches of unknown convergence duration, and with the classical multi-resolution approaches with gradient descents at each scale.

Note also some similarity with recent work on optical flow [14], consisting in an arrangement of 3 different scale-related blocks, though monomodal, not principled from a scale analysis and without scale-specific training.

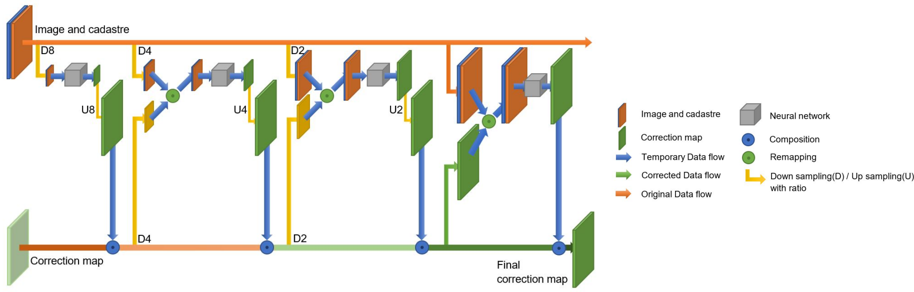


Fig. 4. Full architecture as a chain of scale-specific neural networks. The two full-resolution input images are always available on the top horizontal row. The full-resolution deformation is iteratively refined, scale per scale, on the bottom horizontal line. Each scale-specific block downsamples the images to the right size, applies the previously-estimated deformation, and refines it, in a way somehow similar to ResNet.

We will also check the following variations:

- "*scale-invariant*": replace all scale-specific blocks with the same $s = 2$-specific block, to see how well features generalize across scales; the output quality decreases slightly but remains honorable.
- "*symmetry-invariant*": apply the network on symmetrised and rotated versions of the input images, and average the result over these 8 tests. This ensures rotation invariance and improves the result.

4 Experiments

We perform four experiments on different datasets. The **first experiment** uses the **Inria aerial image labeling dataset** [19], which is a collection of aerial orthorectified color (RGB) imagery with a spatial resolution of 30 cm/pixel covering 810 km^2 over 9 cities in USA and Austria. We aim at aligning a map of buildings downloaded from OSM with the images from the Inria dataset. The network described in Section 3.6 is trained using image patches from six different cities, for which accurate building cadastral data are available[3]. We then evaluated the network by using images of the area of Kitsap County not presented during training. Figure 1 shows an example close-up of alignment result.

In the **second experiment**, the network trained in the first experiment is used to align the OSM building map with satellite images with a pansharpened resolution of 50 cm/pixel, acquired by the Pléiades sensor over the Forez rural area in France. To measure performance of the network, we use the percentage of correct key point metric [29]. We manually identified matching key points on two couples of multimodal images (one Kitsap image from experiment 1 and one Forez image from experiment 2) with more than 600 keypoints for each image.

[3] The cadastral data are extracted from OSM and contain a small misalignment of an order of several pixels.

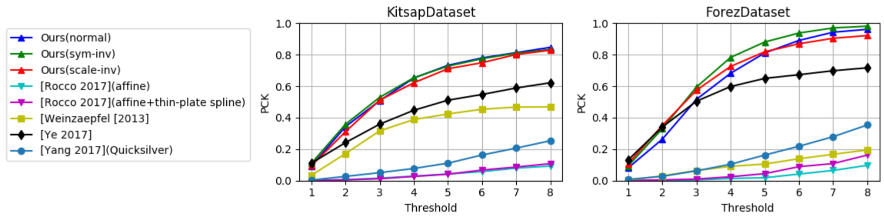
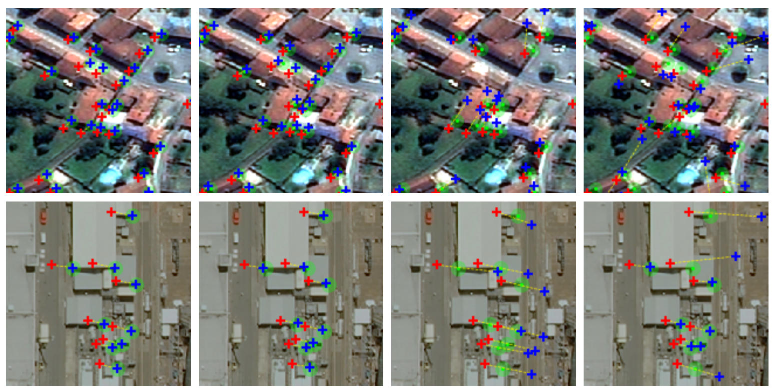


Fig. 5. Key points matching. Scores of different methods on the Kitsap and Forez datasets. The curves indicate the fraction of keypoints whose distance to the ground truth is less than the threshold in pixels. Higher is better.



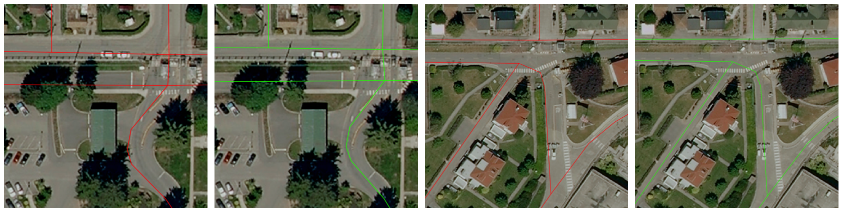
(a) Ground truth (b) Ours (normal) (c) Rocco [29] (d) Weinzaepfel [38]

Fig. 6. Multimodal keypoint matching comparison for different methods and two datasets. Top: Forez dataset (close-up); bottom: Kitsap dataset (close-up). Blue: predicted, green: ground truth. Full resolution in the supplementary materials.

We then measure the distance in pixels between the positions of keypoints after alignment by using different algorithms and the manually indicated ones. If this distance is smaller than a certain threshold, the keypoint is identified as matched. We measure the distance in pixels and not in terms of image proportion because in remote sensing pixels have a ground size in meters regardless of their size. Ultimately we are interested in the alignment error in terms of meters.

Figure 5 compares the performance of our network with the following methods: DeepFlow of Weinzaepfel *et al.* [38], two variations of geometric matching method of Rocco et al. [29], a multimodal registration method of Ye et al. [41], and the deep learning architecture from Yang et al. [39] for medical imagery alignment. Our approach clearly outperforms other ones by a large margin. The reason why neural network approaches in the literature [39,16] do not work on this task is that they were not meant to deal with scale. They were validated on brain registration only, whose typical shifts are of a few pixels (and not 20 or 30 pixels as here). This is coherent with the observations in Section 3.4.

We note that averaging over rotations and symmetries (in green, "*sym-inv*") does help on the Forez dataset, and that learning scale-specific features performs



(a) (b) (c) (d)

Fig. 7. Example of road alignment. (a) and (c): original alignment between image and roads (Kitsap); (b) and (d): results after realignment, respectively.

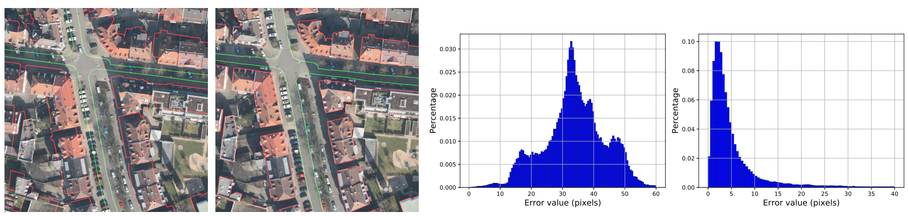


Fig. 8. Example of alignment on the Kitti Dataset. Left: before alignment. Middle: after alignment. Right: misalignment histograms (original misalignment distribution on the top, remaining error on the bottom).

slightly better than scale-independent features but not always (blue vs. red, "*scale-inv*"). Examples of alignment results are shown in Figure 6. Our approach is also much faster than classical approaches, as shown by the computational times below for a 5000×5000 image, even though we compute a dense registration while other approaches only match keypoints:

Method	Ours (normal)	[29]	[38]	[41]
Time	**80 s**	238 s	784 s	9550 s
CPU	Opteron 2Ghz	Intel 2.7Ghz		Int. 3.5Ghz
GPU	GTX 1080 Ti	Q.M2000M		GT 960 M

In a **third experiment**, we align roads with the images used in the first experiment. The task differs from previous experiments in that only the center line of the road is known, and in the form of a polyline. Moreover, local edges are not useful features for alignment anymore, as the center of roads is homogeneous. We train on OSM data, by dilating road polylines to reach a 4 pixel width and rasterizing them. We then test the performance of the trained network on the Kitsap image. The results are shown in Figure 7.

The **fourth experiment** checks the performance of our approach on a higher-resolution dataset. We consider the Kitti dataset [8], which contains high precision aerial images (9 cm/pixel), as well as perfectly aligned multi classes labeling [21]. We create a training set with artificial random deformations, in the same spirit as before, and a test set with randomly deformed images as well, but following different distributions, in order to check also the robustness of our training approach. Image pairs to be registered consist of a RGB image and a 3-channel binary image indicating buildings, roads and sidewalk presence respectively. An example of result is shown in Figure 8. We also analyse the distribution of misalignments before and after registration, shown as histograms in Figure 8. We note that the vast majority of pixels are successfully very closely matched to their ground truth location.

We also perform an extra experiment to show that our multi-scale approach could generalize to other applications. We consider the problem of stereovision, where the input is a pair of RGB images taken from slightly different view points, and the expected output is the depth map, *i.e.* a single channel image instead of a deformation field. We consider the dataset from [23,24] and define the loss function as the depth error (squared), plus the regularizer (4). We keep the same architecture but link the scale-specific networks with additions instead of compositions, so that each block adds scale-specific details to the depth map. The promising result (first run, no parameter tuning) is shown in the supplementary materials, available at https://www.lri.fr/~gcharpia/alignment/ .

Optimization details. The network is trained with an Adam optimizer, on mini-batches of 16 patches of 128×128 pixels images, with a learning rate starting from 0.001 and decayed by 4% every 1000 iterations. Weights are initialized following Xavier Glorot's method. We trained for 60 000 iterations. More technical details are available in the supplementary materials.

Additional details specific to sparse modalities such as cadastral maps, though not essential. During training, we sort out fully or mostly blank images (*e.g.* cadastre without any building). Also, to train more where there is more information to extract (*e.g.*, corners and edges vs. wide homogeneous spaces), we multiply the pixel loss by a factor > 1 on building edges when training.

When rectangular building are glued together in a row with shared walls, the location of their edges and corners is not visible anymore on the rasterized version of the OSM cadastre. By adding a channel to the cadastre map, reminding the OSM corner locations, we observe a better alignment of such rows.

5 Conclusion

Based on an analysis of classical methods, we designed a chain of scale-specific neural networks for non-rigid image registration. By predicting directly the final registration at each scale, we avoid slow iterative processes such as gradient descent schemes. The computational complexity is linear in the image size, and far lower than even keypoint matching approaches. We demonstrated its performance on various remote sensing tasks and resolutions. The trained network as well as the training code will be made available online. This way, we hope to contribute to the creation of large datasets in remote sensing, where precision so far was an issue requiring hand-made ground truth.

Acknowledgements. This work benefited from the support of the project EPITOME ANR-17-CE23-0009 of the French National Research Agency (ANR).

References

1. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision **61**(2), 139–157 (2005)
2. Bischke, B., Helber, P., Folz, J., Borth, D., Dengel, A.: Multi-task learning for segmentation of building footprints with deep neural networks. arXiv preprint arXiv:1709.05932 (2017)
3. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a" siamese" time delay neural network. In: Advances in Neural Information Processing Systems. pp. 737–744 (1994)
4. Charpiat, G., Keriven, R., Faugeras, O.: Image statistics based on diffeomorphic matching. In: ICCV'05. vol. 1, pp. 852–857
5. Charpiat, G., Maurel, P., Pons, J.P., Keriven, R., Faugeras, O.: Generalized gradients: Priors on minimization flows. International Journal of Computer Vision (2007)
6. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR'05. vol. 1, pp. 539–546
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR'05. vol. 1, pp. 886–893. https://doi.org/10.1109/CVPR.2005.177
8. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
9. Glaunes, J., Trouvé, A., Younes, L.: Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. In: CVPR'04. vol. 2, pp. II–II. Ieee
10. Haklay, M., Weber, P.: Openstreetmap: User-generated street maps. IEEE Pervasive Computing **7**(4), 12–18 (Oct 2008)
11. Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G.: High-resolution global maps of 21st-century forest cover change. Science **342**(6160), 850–853 (2013)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
13. Hermosillo, G., Chefd'Hotel, C., Faugeras, O.: Variational methods for multimodal image matching. International Journal of Computer Vision **50**(3), 329–343 (2002)
14. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. arXiv preprint arXiv:1612.01925 (2016)
15. Kendall, D.G.: A survey of the statistical theory of shape. Statistical Science pp. 87–99 (1989)
16. Kwitt, R., Niethammer, M.: Fast predictive simple geodesic regression. In: third International Workshop DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017. vol. 10553, p. 267. Springer (2017)
17. Lee, D., Hofmann, M., Steinke, F., Altun, Y., Cahill, N.D., Scholkopf, B.: Learning similarity measure for multi-modal 3d image registration. In: CVPR'09. pp. 186–193
18. Maggiori, E., Charpiat, G., Tarabalka, Y., Alliez, P.: Recurrent neural networks to correct satellite image classification maps. IEEE Transactions on Geoscience and Remote Sensing **55**(9), 4962–4971 (Sept 2017). https://doi.org/10.1109/TGRS.2017.2697453

19. Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P.: Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In: IGARSS'07
20. Maintz, J.B.A., van den Elsen, P.A., Viergever, M.A.: Evaluation of ridge seeking operators for multimodality medical image matching. IEEE Transactions on pattern analysis and machine intelligence **18**(4), 353–365 (1996)
21. Máttyus, G., Wang, S., Fidler, S., Urtasun, R.: Hd maps: Fine-grained road segmentation by parsing ground and aerial images. In: CVPR'16. pp. 3611–3619
22. Meinhardt, T., Möller, M., Hazirbas, C., Cremers, D.: Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In: ICCV'17
23. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR'15
24. Menze, M., Heipke, C., Geiger, A.: Joint 3d estimation of vehicles and scene flow. In: ISPRS Workshop on Image Sequence Analysis (ISA) (2015)
25. Merkle, N., Luo, W., Auer, S., Mller, R., Urtasun, R.: Exploiting deep matching and sar data for the geo-localization accuracy improvement of optical satellite images. Remote Sensing **9**(6) (2017). https://doi.org/10.3390/rs9060586, http://www.mdpi.com/2072-4292/9/6/586
26. Michor, P.W., Mumford, D., Shah, J., Younes, L.: A metric on shape space with explicit geodesics. arXiv preprint arXiv:0706.4299 (2007)
27. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International journal of computer vision **60**(1), 63–86 (2004)
28. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: CVPR'17. vol. 2
29. Rocco, I., Arandjelovic, R., Sivic, J.: Convolutional neural network architecture for geometric matching. CoRR **abs/1703.05593** (2017), http://arxiv.org/abs/1703.05593
30. Rohé, M.M., Datar, M., Heimann, T., Sermesant, M., Pennec, X.: Svf-net: Learning deformable image registration using shape matching. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 266–274. Springer (2017)
31. Schnabel, J.A., Rueckert, D., Quist, M., Blackall, J.M., Castellano-Smith, A.D., Hartkens, T., Penney, G.P., Hall, W.A., Liu, H., Truwit, C.L., et al.: A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 573–581. Springer (2001)
32. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM International Conference on Multimedia. pp. 357–360. MM '07, ACM, New York, NY, USA (2007). https://doi.org/10.1145/1291233.1291311, http://doi.acm.org/10.1145/1291233.1291311
33. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. IEEE transactions on medical imaging **32**(7), 1153–1190 (2013)
34. Sundaramoorthi, G., Yezzi, A., Mennucci, A.: Coarse-to-fine segmentation and tracking using sobolev active contours. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(5), 851–864 (2008)
35. Verdie, Y., Lafarge, F.: Efficient Monte Carlo sampler for detecting parametric objects in large scenes. In: ECCV'12
36. Von Eicken, T., Basu, A., Buch, V., Vogels, W.: U-net: A user-level network interface for parallel and distributed computing. In: ACM SIGOPS Operating Systems Review. vol. 29, pp. 40–53. ACM (1995)

37. Wang, S., Fidler, S., Urtasun, R.: Proximal deep structured models. In: Advances in Neural Information Processing Systems. pp. 865–873 (2016)
38. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: DeepFlow: Large displacement optical flow with deep matching. In: ICCV (2013), http://hal.inria.fr/hal-00873592
39. Yang, X., Kwitt, R., Niethammer, M.: Quicksilver: Fast predictive image registration - a deep learning approach. CoRR **abs/1703.10908** (2017)
40. Ye, Y., Shan, J.: A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences. ISPRS Journal of Photogrammetry and Remote Sensing **90**, 83–95 (2014)
41. Ye, Y., Shan, J., Bruzzone, L., Shen, L.: Robust registration of multimodal remote sensing images based on structural similarity. IEEE Transactions on Geoscience and Remote Sensing **55**(5), 2941–2958 (2017)
42. Ye, Y., Shen, L.: Hopc: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. In: Proc. Ann. Photogram., Remote Sens. Spatial Inf. Sci.(ISPRS). pp. 9–16 (2016)
43. Yu, L., Zhang, D., Holden, E.J.: A fast and fully automatic registration approach based on point features for multi-source remote-sensing images. Computers & Geosciences **34**(7), 838–848 (2008)
44. Zbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. CoRR **abs/1510.05970** (2015), http://arxiv.org/abs/1510.05970