3D reconstruction of curvilinear structures with stereo matching deep convolutional neural networks

Okan Altingövde [a,*], Anastasiia Mishchuk [a,b], Gulnaz Ganeeva [b], Emad Oveisi [c], Cecile Hebert [b], Pascal Fua [a]

[a] *Computer Vision Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland*
[b] *Electron Spectrometry and Microscopy Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland*
[c] *Interdisciplinary Centre for Electron Microscopy, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland*

ARTICLE INFO

ABSTRACT

Curvilinear structures frequently appear in microscopy imaging as the object of interest. Crystallographic defects, i.e dislocations, are one of the curvilinear structures that have been repeatedly investigated under transmission electron microscopy (TEM) and their 3D structural information is of great importance for understanding the properties of materials. 3D information of dislocations is often obtained by tomography which is a cumbersome process since it is required to acquire many images with different tilt angles and similar imaging conditions. Although, alternative stereoscopy methods lower the number of required images to two, they still require human intervention and shape priors for accurate 3D estimation. We propose a fully automated pipeline for both detection and matching of curvilinear structures in stereo pairs by utilizing deep convolutional neural networks (CNNs) without making any prior assumption on 3D shapes. In this work, we mainly focus on 3D reconstruction of dislocations from stereo pairs of TEM images.

1. Introduction

Transmission Electron Microscopy delivers a 2D projection of the sample under consideration while many areas of biological and physical sciences now require 3D reconstructions of nano-sized objects. In particular, curvilinear objects such as dislocations [1,2] are of great interest in material science.

The traditional approach to recovering the third dimension is to tilt the sample with respect to the electron beam, to capture dozens of views, and to perform tomographic reconstruction [3–7]. This is time consuming and usually delivers a data volume from which the dislocations still have to be traced manually. An alternative is to use stereography instead. It only requires a few images. Unfortunately, state-of-the-art methods [8–11] still involve substantial manual intervention for line detection and matching.

In this paper, we present a fully automated stereo-based approach that only requires two images and directly outputs the 3D reconstructions in usable form. Its accuracy nevertheless is similar to that of tomography. Fig. 1 depicts our proposed pipeline. Its key component is a deep network that jointly detects dislocations in individual images and matches them across images to recover their 3D structure, while being end-to-end trainable in a semi-supervised way. The latter is important because ground truth is rarely, if ever, available. A key

improvement over earlier methods is that the detection and matching steps are performed jointly instead of sequentially. In addition to eliminating manual intervention, it stops the algorithm from finding spurious dislocations.

Depth information can be recovered from two or more images by triangulating the 2D projections of specific 3D points that have been matched across images. Such points are known as feature-points. They can either be sparsely sampled across the images or form linear structures.

In the early days of computer vision, computational power was very limited and such feature-based stereo-reconstruction techniques arose from the need to limit the computational requirements by operating only on a small fraction of the image. An additional strength of these algorithms was that, because features could be detected to subpixel-accuracy, good reconstruction accuracy could be achieved. For example, as early as in 1988, the system of [13] achieved precisions in the order of half a pixel in disparity, which has not been improved much since [14]. Unfortunately, it only matched a very small proportion, typically less than 1%, of the image points and in relatively narrow baseline stereo pairs. Among these approaches, contour-based ones became popular in the 1980s [15–17] because they made it
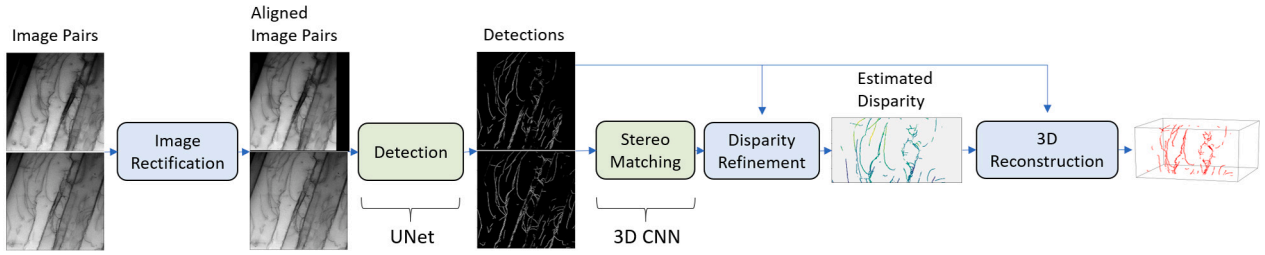

Fig. 1. Stereo reconstruction processing steps. In the first step, images are aligned such that the tilt axis is approximately vertical and at the image center. Dislocation segments are detected using UNet [12] in the second stage. Later, features of UNet are provided to a 3D CNN in order to compute horizontal disparity between corresponding dislocation segments. At the final stage, depth value for every detected dislocation pixel is computed using disparity estimations. Green blocks in the depicted figure represent the stages that are learned from annotated training data.
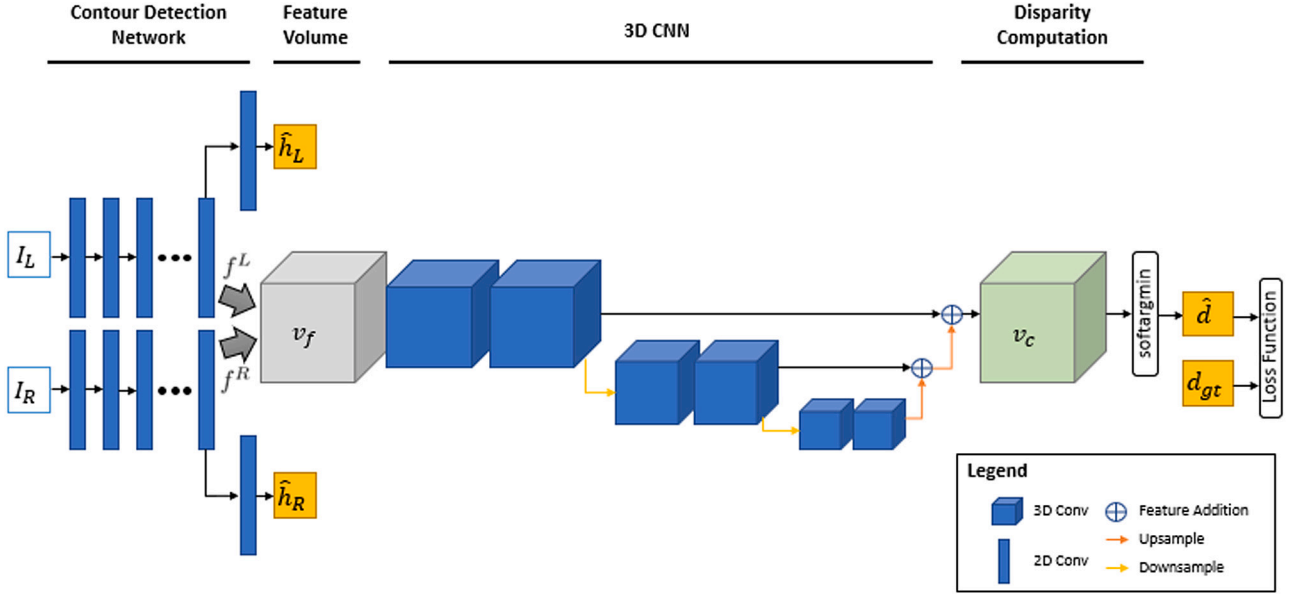

Fig. 2. Stereo Contour Detection and Matching Network architecture. Layers with learnable parameters are shown in blue. Feature volume, V_f, and cost volume, V_c, are depicted as gray and green 3D volumes respectively.

possible to reason about contour continuity and impose additional geometric constraints. However, with the advent of algorithms based on correlation [18], graph-cut [19], or deep-learning [20–23] that could match much larger fractions of the images at a computational cost that modern hardware can easily handle, all these feature-based techniques have fallen out of favor and there has been remarkably few new developments in the last 20 years.

Following the advancements in hardware and deep learning research, it has become the norm to compute dense features by applying convolutions to each pixel location. [24,25] are typical network architectures that have been widely employed as feature extraction sub-network. The network architecture proposed in [12] has been successful specifically in microscopy field due its ability to efficiently fuse multi-scale features. Hence, we deploy it as a pre-processing sub-network.

Our proposed pipeline delivers reconstruction accuracy comparable to that of tomography while reducing the imaging and processing times. Moreover, it automatically detects dislocations and saves us from filtering the final reconstruction with manually set thresholds.

2. Method

Conceptually, our stereo reconstruction pipeline follows the classical paradigm and comprises the four following steps: Image rectification, detection of the linear structures, stereo matching, and finally triangulation to turn the 2D contours into 3D dislocation models. These

fours steps are depicted by Fig. 1. In the earlier era described above, they would have been performed sequentially and independently from each other. However, with the advent of Deep Learning [26], it has become both possible and desirable to merge detection and matching and to have them be performed jointly by the end-to-end trainable deep net, such as the one shown in Fig. 2.

In the remainder of this section, we first discuss our approach to detection and matching of dislocations in 2D images in the ideal case where these images are properly aligned. We then discuss how they can be *rectified* so that they become aligned and finally how to generate fully 3D dislocations from the 2D projections.

2.1. Detecting and matching dislocations

In two-view stereo, given a point in the first image, recovering its 3D position involves finding the corresponding point along the so-called *epipolar* line. If the two images are coplanar, that is, they were taken such that the second camera is only offset horizontally compared to the first one, then each pixel's epipolar line is horizontal and at the same vertical position as that pixel. This also applies to pairs of TEM images because they are orthographic projections of the same object on a fixed detector/camera acquired while the object is being tilted in the microscope.

In this section, we will assume this to be the case, that is, for a point, $p = (x, y)$ in the first image, the corresponding point q is at location $(x + d, y)$ in the second image, where d is the *disparity* and
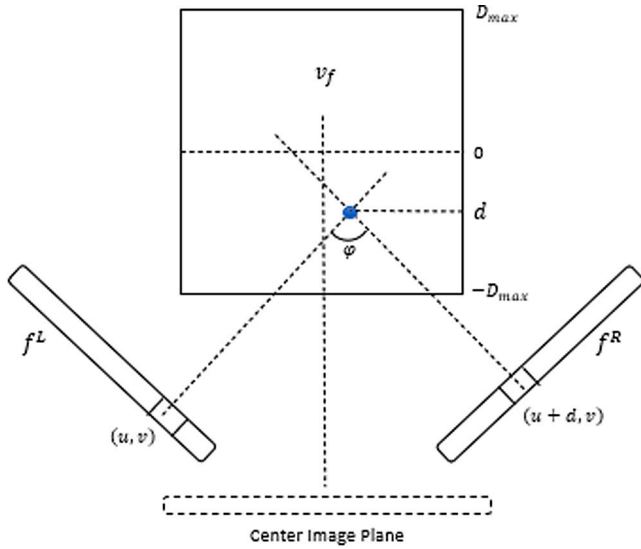
Fig. 3. Feature volume construction is shown from top view. For each pixel in left image, features of this pixel and candidate pixels for a match in the right image are back-projected to 3D and concatenated to form feature volume. For a specific pixel at location (u, v) in left image and a candidate pixel at location $(u + d, v)$, back-projected point in feature volume is depicted as blue circle in the figure. The value d for the given pixel is the disparity level that defines the depth of the pixel with respect to the center image plane.

proportional to the distance to the image plane. We will show in the following section how the images can be rectified to make this be true.

2.1.1. Network architecture

To detect the dislocations in the 2D images and compute disparities for each one of their points, we use the deep network depicted by Fig. 2. Its first stage consists of a 2D *contour detection network* C in a Siamese configuration [27]. The feature maps produced by C for both views are combined into the *feature volume*, which is then refined by a 3D CNN into a *cost volume*.

We describe each building block of architecture in more detail below.

Pre-trained contour detection network. C extracts contour-specific features. We denote the feature maps by f^L and f^R and contour detections by $\hat{h}^L$ and $\hat{h}^R$ for the left and right images, I_L and I_R, respectively. C is trained with binary cross-entropy loss. In Fig. 4, we show an example detection output of C.

Feature volume. As in many existing architectures [28], we construct a feature volume V_f shown in Fig. 3, by replicating the left and right feature maps once for every possible disparity value, and concatenating them into a tensor of size $2 \cdot featuresize \times height \times width \times disparity\ range$, so that $V_f(\cdot, u, v, d)$ is a concatenation of features $f^L(\cdot, u, v)$ and $f^R(\cdot, u + d, v)$, where u and v are coordinates of the left image and d is disparity. This enables the evaluation of match hypotheses (u, v, d) by means of a 3D CNN as follows.

3D CNN For matching cost estimation. A 3D CNN is used to compute a cost volume V_c, shown in green in Fig. 2, from the feature volume V_f. V_c is of size $height \times width \times disparity\ range$ and an entry $V_c(u, v, d)$ encodes the cost of matching a feature f^L at position u, v in the left feature map, to a feature $f^R(u + d, v)$ of the right feature map. 3D CNN consists of multiple stacked hourglass modules composed of 3D $3 \times 3 \times 3$ convolution layers, shown in blue volumes in the center of Fig. 2.

Disparity computation. Optimum disparity for a pixel at location (u, v) is the one with lowest matching cost which is encoded in cost volume, V_c. Therefore, a disparity estimate is the index of the minimum element along disparity axis of V_c for each pixel, which could be found using the *argmin* operation. However, *argmin* is not differentiable and can therefore not be used as part of neural network architecture. Instead we use the differentiable *soft-argmin* defined as

$$\hat{d}_{uv} = \sum_{d=0}^{d_{max}} d \cdot \sigma(-V_c(u, v, d)) \tag{1}$$

where $\sigma(.)$ is the *softmax* operation, d is disparity level and $V_c(u, v, d)$ is matching cost of pixel at location (u, v) for disparity d to be able to train our network. We refer interested readers to [28] for further information on *soft-argmin* function.

2.1.2. Loss function

We train our stereo matching network with a loss function that is a weighted sum of three loss terms

$$L = \gamma_{disp}L_{disp} + \gamma_{var}L_{var} + \gamma_{warp}L_{warp}, \tag{2}$$

where L_{disp}, L_{var}, and L_{warp} represent a disparity, variance, and warp loss, which we describe in more detail below.

Disparity Loss. When the ground-truth disparity, d_{gt}, is available, we penalize inconsistencies in the prediction $\hat{d}$ by means of the smooth L1 loss

$$L_{disp} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \begin{cases} 0.5 \cdot e_d{}^2, & \text{if } e_d < 1 \\ e_d - 0.5, & \text{otherwise} \end{cases}, \tag{3}$$

where $e_d = \left\| \hat{d} - d_{gt} \right\|$ and $\mathcal{A}$ is the set of pixels belonging to dislocations with available disparity annotations (see Fig. 5).

Variance Loss. While minimizing L_{disp} encourages the network to find the correct disparity value, it is also important that it delivers a low variance around that ground truth value so that the estimates are accurate. Therefore, we define variance loss

$$L_{var} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \hat{v}, \tag{4}$$

where $\hat{v}_{uv} = \sum_{d=0}^{d_{max}} (d - d_{gt})^2 \cdot \sigma(-V_c(u, v, d))$.

Disparity estimator defined in Eq. (1) may be seen as probability-average of possible disparity levels by taking $\sigma(-V_c(u, v, d))$ term as probability of disparity level d. In this case, $\hat{v}$ becomes expectation of the squared deviations of disparity around the disparity ground truth. Minimizing L_{var} therefore minimizes the variance of the estimate $\hat{v}$ around d_{gt}.

Warp Loss. Given the detection map of curvilinear structures in the right image, $\hat{h}^R$ and the disparity, $\hat{d}$, the left detection map, $\hat{h}^L$ can be reconstructed by shifting pixels with their corresponding disparity values. This operation is called *warping* and we incorporate it in our warp loss

$$L_{warp} = L_{bce}(warp(\hat{h}^R, \hat{d}), \ h_{gt}^L). \tag{5}$$

where L_{bce} is the binary cross entropy loss. Minimizing L_{warp} enforces consistency between $\hat{d}$, $\hat{h}^R$ and $\hat{h}^L$.

Overall performance of 3D CNN depends on its capability to localize minimum cost disparity level in V_c for each pixel. In other words, it is required to have low variance probability distributions over estimated disparity values to confidently make predictions. In this sense, complementing disparity loss with variance and warp loss terms allows us to further constrain network training to have cost estimations not only centered at ground truth level but also having lower variances around it. This is especially important when training data is not abundant and cumbersome to annotate which is the case for electron microscopy imaging. In result section, we present an ablation study on components of loss function to show strong evidence in favor of using complemented loss function over naive disparity loss.

(a)

(b)
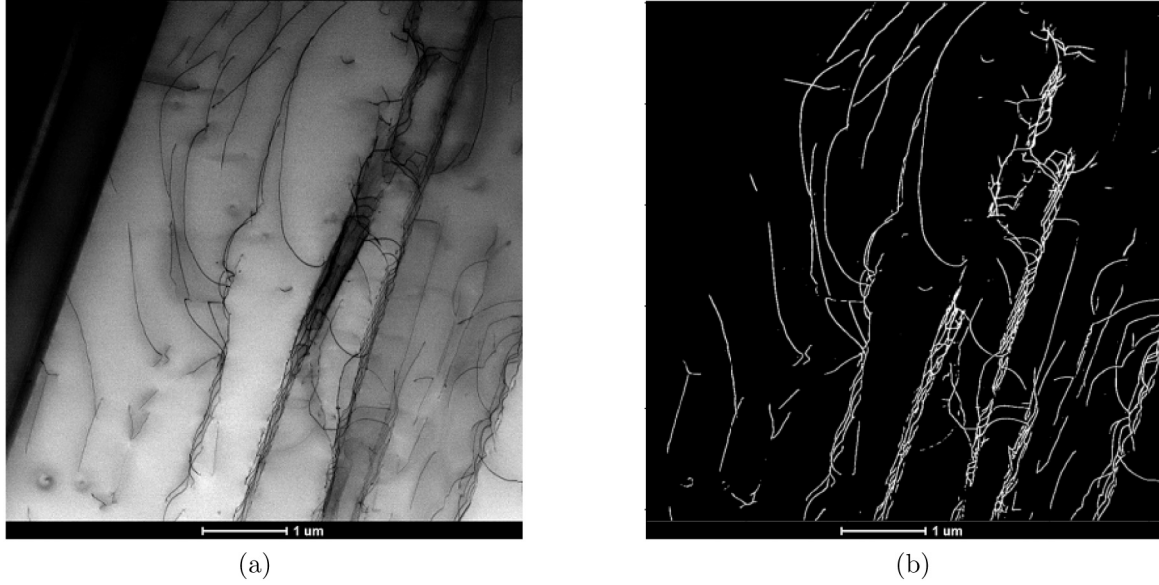
Fig. 4. Representative 2D ABF STEM image of dislocations in a TiAl super-alloy thin foil in two-beam diffraction condition with **g** = (002) (a) and corresponding detected dislocations (b). It might be seen that detection network manages to detect dislocations having complex shapes in noisy images while still being able to reject unrelated dark sections of the image.
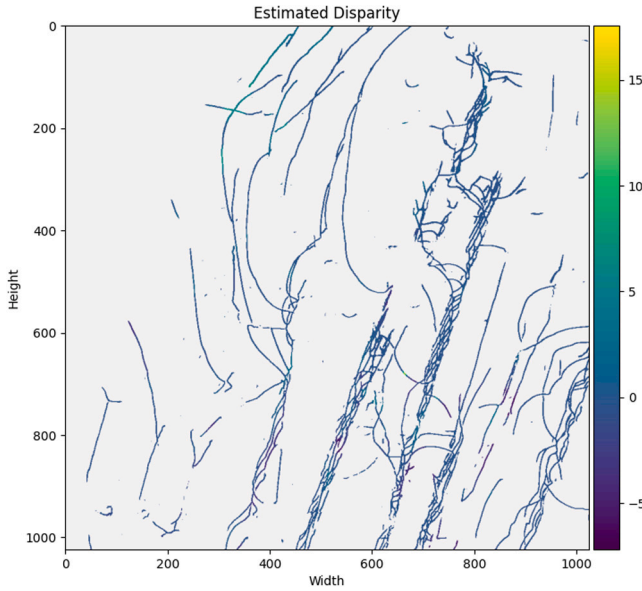
Fig. 5. Estimated disparity of detected dislocations shown in Fig. 4. The disparity map is made using pair of stereo images with 4 degrees of tilt between two images. Depth axis is towards image plane. Dislocations segments with bright colors are closer to the image plane. Width, height and disparity values are all in pixels. Best viewed in color.

2.2. Image rectification

In practice, the images are never perfectly aligned. Although the viewing angle can be controlled up to a degree, the region of interest is not necessarily located around the rotation axis and need to be centered manually during image acquisition. Manual search of the region causes translational misalignment between consecutive images acquired. Moreover, it is not always possible to maintain exact vertical tilt-axis due to a non-perfectly compensated image rotation which may yield deviations up to 2° from the vertical axis. Thus, we write the relationship between image coordinates of corresponding points $\langle p, q \rangle$

in the original unaligned images as

$$(x_p, y_p) = (x_q + d_q \cos \theta - t_x, y_q + d_q \sin \theta - t_y),$$
$$(\Delta x_{pq}, \Delta y_{pq}) = (d_q \cos \theta - t_x, d_q \sin \theta - t_y), \tag{6}$$

where θ is the angle the tilt-axis makes with the vertical axis on the image plane, d_q is disparity of point q, (t_x, t_y) are translational shifts that occurred during acquisition on axes x and y of image plane respectively and $(\Delta x_{pq}, \Delta y_{pq})$ are total displacements on image plane between correspondences prior to alignment.

Hence the image transformation that corrects this misalignment is

$$T = [I \mid t]$$
$$(\hat{x}_q, \hat{y}_q)^T = T_{[2 \times 3]}(x_q, y_q, 1)^T \tag{7}$$

where I is the 2×2 identity matrix and T is the 2×3 transformation matrix that translates point coordinates expressed in projective coordinates.

We automatically detect and match Scale Invariant Feature Transform (SIFT) [29] keypoints to construct sparse correspondences between left and right images. These points are used to estimate the transformation T by iteratively solving a linear system of equations. At each iteration, outliers are eliminated using random sample consensus (RANSAC) [30]. After applying T to the right image, the total displacement between correspondences, $(\Delta x_{p\hat{q}}, \Delta y_{p\hat{q}})$ becomes $(d_q \cos\theta, d_q \sin\theta)$ and θ can be computed as

$$\theta = \arctan(\frac{\Delta y_{p\hat{q}}}{\Delta x_{p\hat{q}}}). \tag{8}$$

Finally, the images are aligned by applying the rotation matrix

$$R = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \tag{9}$$

to the transformed right image. Once this is done, the epipolar lines are parallel to the horizontal axis of the images, which is a prerequisite for inputs to our deep network as discussed in the beginning of this section.

2.3. Disparity refinement

Disparity values are computed from the estimated cost volume using a soft-argmin operation as shown in Eq. (1). In practice, this yields
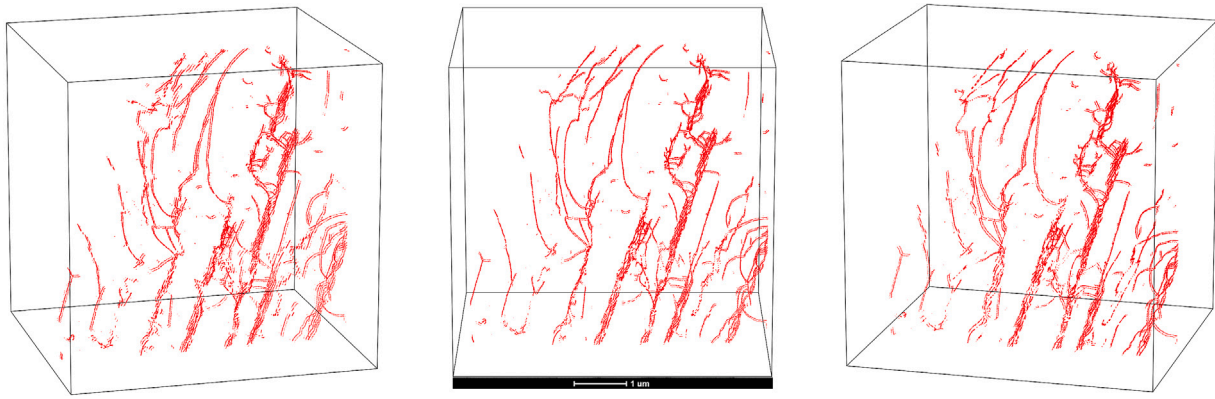
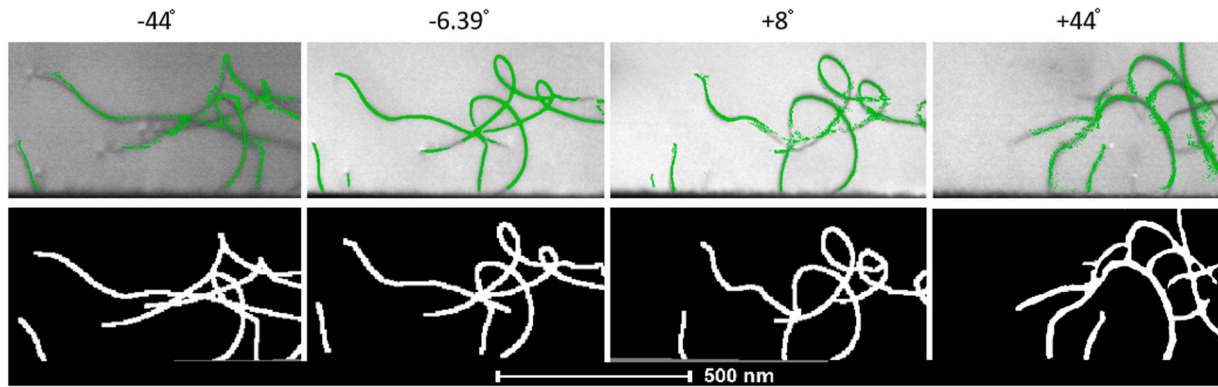**Fig. 6.** 3D Reconstruction of dislocations shown in Fig. 4.

Fig. 7. Re-projections to other views. Estimated 3D dislocations in TiAl sample that are reconstructed using a stereo pair with 8 degree separation are re-projected to views with −44, −6.39, +8 and +44 degrees tilt angles. In the top row of images, four views are shown with green marks being 3D dislocation points re-projected onto raw images. In the bottom row, dislocation ground truth for these views are shown. For all presented views, it can be seen that re-projections of the estimated 3D structure are consistent with raw images and human annotated dislocations ground truth.

disparities that are within one or two pixels of what they should be. To improve accuracy, we use a standard technique to refine them [31,32]: We use them to warp the dislocations found in the left image into estimates in the right image, which we then match to the closest dislocation actually found in the right image to refine the disparity value. Warped dislocations are matched pixel-wise to the closest candidate within a 3 pixel distance, which delivers the pixel accuracy required for accurate reconstruction. Dislocation pixels without any suitable candidate in this range are marked as erroneous and removed from the final 3D reconstruction.

In the result section, we will compare reconstructions obtained with and without this refinement stage and show that, on average, it improves precision.

2.4. 3D Reconstruction

Having disparity values computed, the 3D shapes of curvilines can be reconstructed by applying triangulation with given viewing angles and the orthographic camera assumption.

In our stereo setup of verged orthographic cameras which is shown in Fig. 3, we define the depth axis same as disparity axis which is parallel to surface normal of the imaginary center image. Given this definition of depth, its relation to the disparity d is

$$\text{depth} = \frac{d}{2sin(\frac{\phi}{2})} , \tag{10}$$

where ϕ is the angle between two views. The term $sin(\frac{\phi}{2})$ in Eq. (10) acts as a scaling factor between depth and disparity. In Fig. 6, a 3D reconstruction of dislocations from TEM images with 4 degrees between images is shown.

## 3. Experiments

3.1. Datasets

For our experiments, we created three stereo dislocation dataset each composed of stereo pairs taken from image sequences of three different materials, representing large diversity of dislocation shapes and contrasts. Bright-field images of dislocations were acquired from TiAl alloy, GaN, and high-entropy (Cantor) alloy using transmission electron microscope in scanning mode (STEM). Details of imaging conditions and specimen preparation can be found in Appendix.

Our dislocation dataset consists of 90 labeled stereo pairs grayscale images with 512x512 pixels resolution split into train, validation and two test subsets. Segmentation masks and disparity maps were manually labeled for the task. The dataset will be publicly available at www.epfl.ch/labs/cvlab/data/data-dislocations.

3.2. Results

We propose three different qualitative ways to evaluate our results and demonstrate their accuracy. We also present an ablation study for different stereo setups with varying tilt angles and loss functions to quantitatively show their effect on the overall reconstruction performance.

Our method reconstructs 3D shapes of dislocations for images with 1024x1024 resolution approximately in three seconds when it runs on NVIDIA RTX 2070 GPU. Same process takes approximately one minute processing time on modern CPUs. In our experiments, we evaluate detection and matching performance on 512x512 image crops.

5
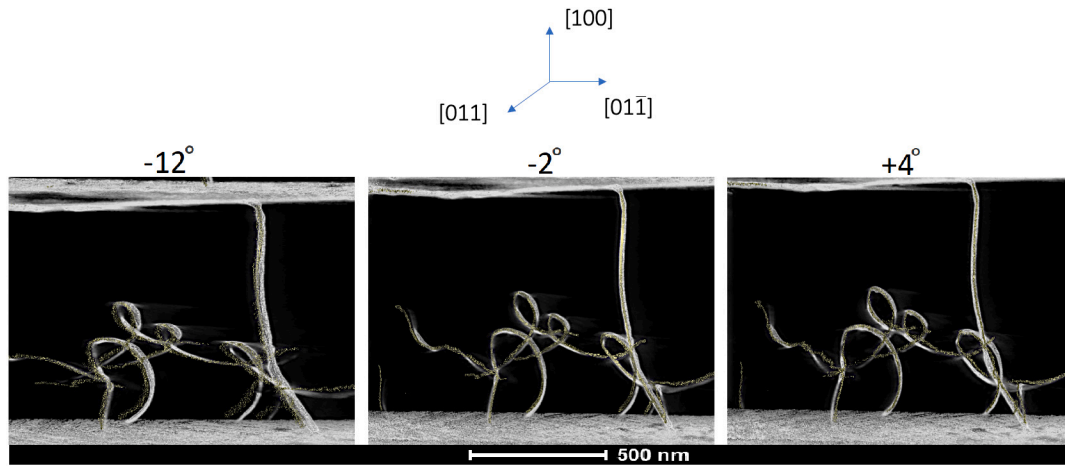
Fig. 8. Comparison of reconstruction results of our approach and tomography. TiAl sample region previously shown in Fig. 7 is reconstructed by tomography (shown in white) utilizing full image sequence with 45 images covering the angular range (−48°, 50°) and by our method with stereo images having 2 degrees in between (shown in yellow).
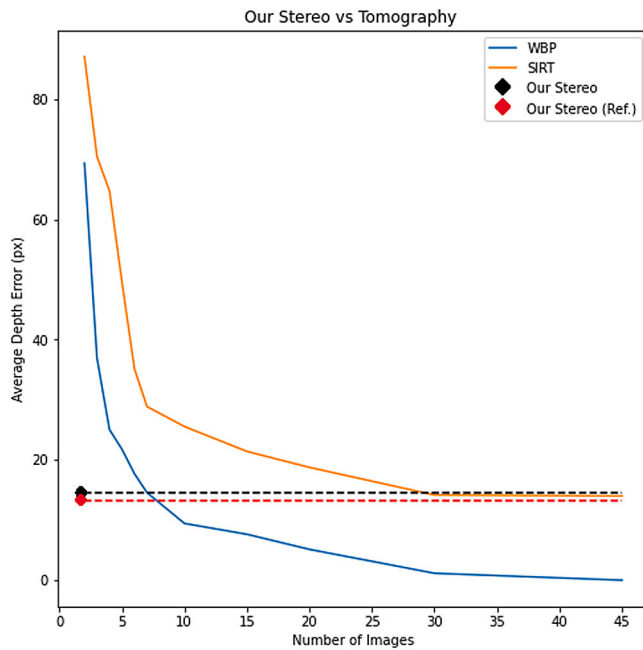
Fig. 9. Two 3D tomography reconstruction techniques are compared to our stereo method across increasing number of images given as input. Images are sampled regularly for each case from the sequence with 45 images spanning angular range (−48°, 50°).

Fig. 10. Samples from our dataset.

Table 1

Comparison to stereo matching baselines. All reported values are mean absolute errors in pixels.

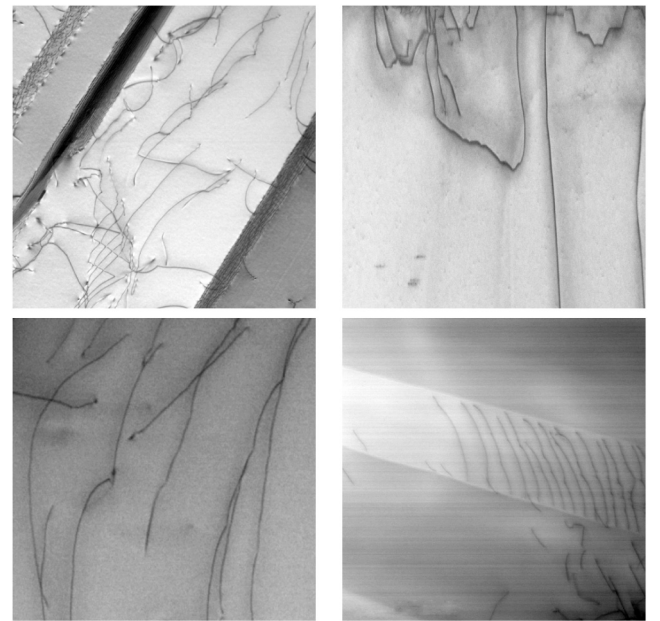| Method/Material	GaN	TiAl	Cantor alloy
Lucas–Kanade [33]	15.31	16.65	17.96
SGM [34]	10.21	11.42	13.21
GC-Net [28]	2.30	3.12	2.15
Ours	1.11	1.41	0.98

3.2.1. Re-projection errors

Once the 3D dislocations have been reconstructed, we can project them into views that were *not* used to perform the reconstruction. When these projections superpose well with the actual dislocations, it is evidence that the 3D reconstructions were correct. In Fig. 7, it is shown for the case where stereo pairs have 8 degrees between them.

3.2.2. Comparison to stereo matching baselines

We compare our method's stereo matching performance to two classical [33,34] and one learning-based [28] stereo matching baselines (see Table 1).

It is worth noting that our method detects and match structures jointly unlike the reported stereo matching baselines. Direct integration of detection to stereo matching improves overall 3D reconstruction significantly. In addition, previous stereo matching methods applied to curvilinear structures [11] relies on human intervention for both detecting and matching the keypoints on structures. To our knowledge, our approach is the first fully automated joint detection and stereo matching approach for curvilinear structures.

3.2.3. Comparison to tomography

To further evaluate the results Fig. 8 shows qualitative comparison of reconstructed 3D dislocations from tomography baseline method and our neural network reconstruction. Tomogram was obtained with Inspect 3D Software using Simultaneous Iterations Reconstruction Technique (SIRT). 3D visualization of the tomogram was performed in
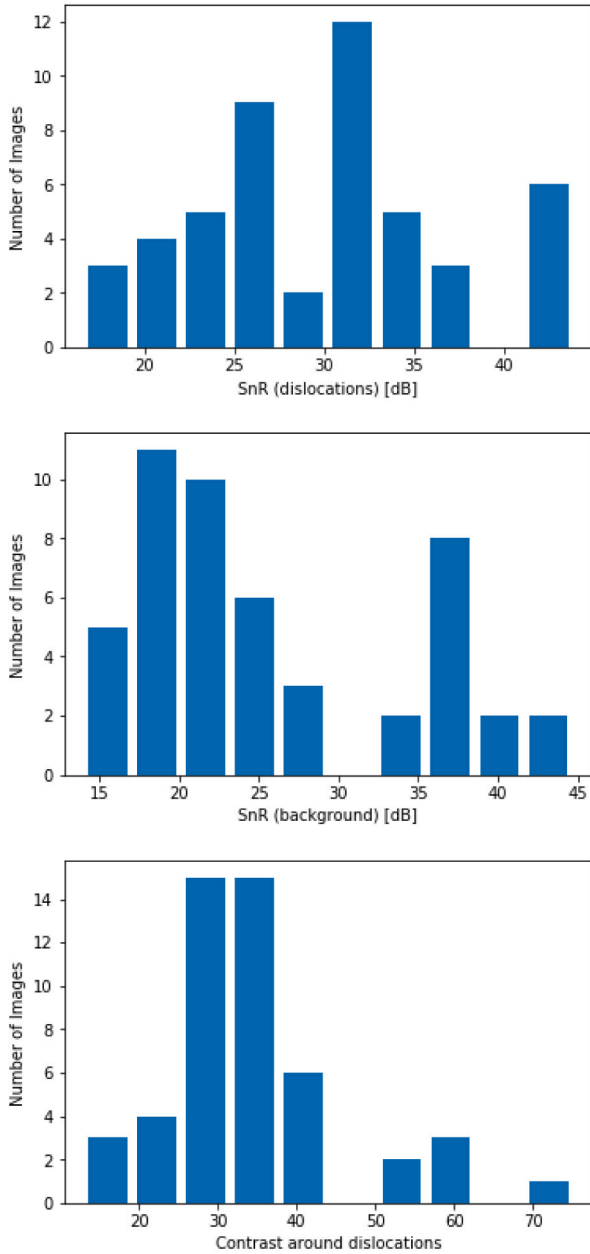
Fig. 11. Training dataset distribution with respect to signal-to-noise ratio and contrast measurements.

Table 2

Depth estimation errors (in pixels) for stereo setups with different angle separations. For all experiments, we trained the network using the loss function that combines all three loss terms. We present results without the refinement stage of Section 2.3 and with it.

	2°	4°	6°	8°	10°	12°
Ours (Raw)	18.3	21.5	15.0	14.6	22.2	28.4
Ours (Refined)	18.0	20.1	14.7	**13.2**	21.8	29.9

Table 3

Comparison of disparity estimation with different combinations of loss terms. Results are presented for three subsets of test data. Stereo pairs with 8° angular range is used for the experiments.

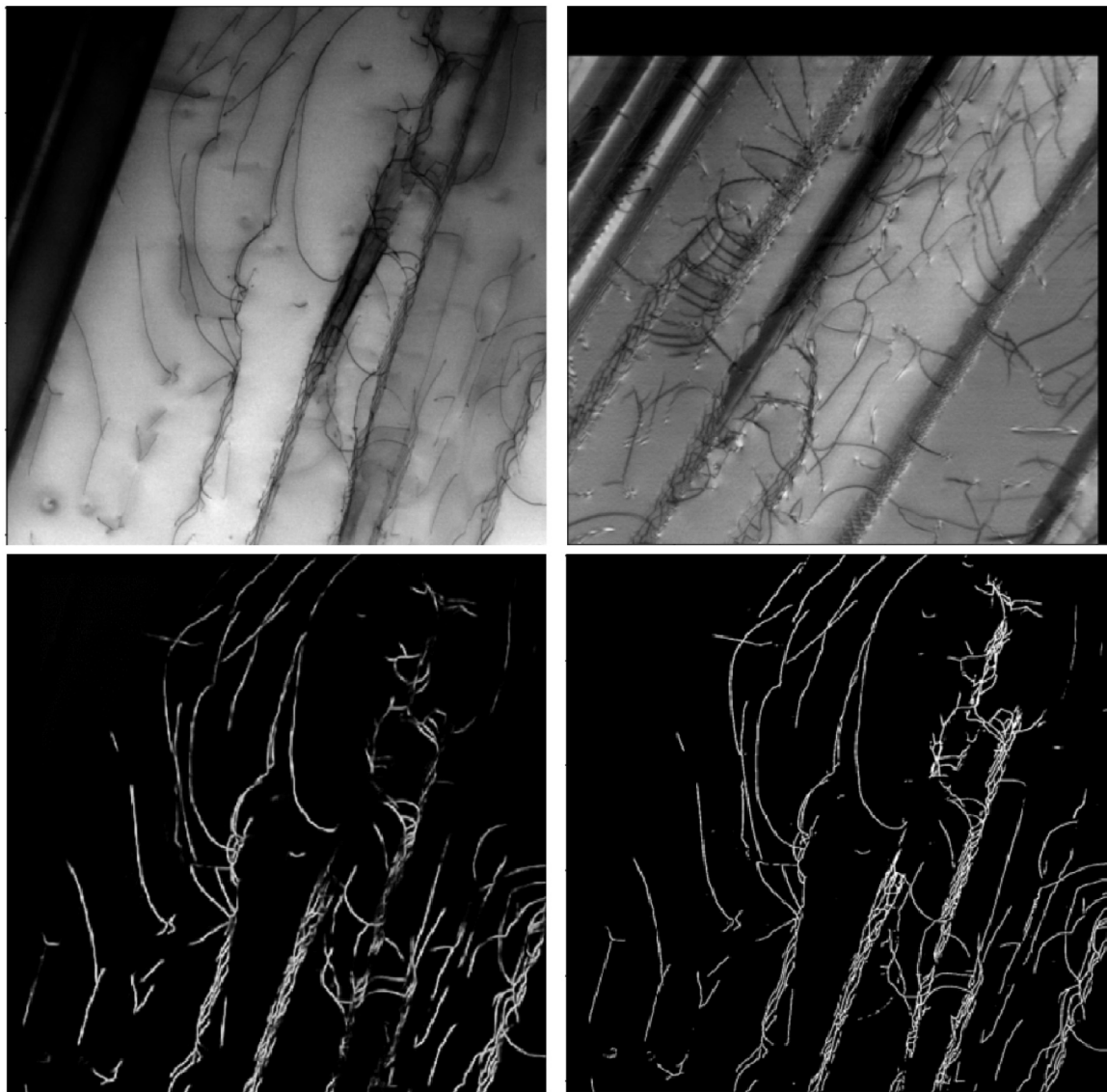| Loss function	1px %	3px %	5px %	EPE px–nm
GaN				
Disp+Warp+Var	**40.1**	**8.3**	**3.5**	**1.112–9.6**
Disp+Var	40.3	8.5	**3.5**	1.120–9.7
Disp+Warp	41	8.6	3.8	1.238–10.7
Disp	52.1	9.4	4.4	1.852–16.0
Warp	81.6	55.3	30.2	5.034–43.6
TiAl				
Disp+Warp+Var	41	**9.6**	**4.1**	**1.411–30.3**
Disp+Var	**40.7**	9.7	4.2	1.418–30.5
Disp+Warp	41.6	9.7	4.2	1.415–30.4
Disp	54.9	14.2	5.8	1.755–37.7
Warp	84.8	57.9	38.3	5.879–126.4
Cantor alloy				
Disp+Warp+Var	**38.2**	**8.0**	**3.1**	**0.982–9.6**
Disp+Var	40.5	8.4	3.3	1.101–10.8
Disp+Warp	41.3	8.7	3.6	1.322–12.9
Disp	49.8	9.0	4.3	1.523–14.9
Warp	80.4	52.4	28.5	5.011–49.2

the only fully-automated stereo reconstruction method for curvilinear structures. We further discuss advantages and drawbacks of the proposed method in discussion section.

3.2.4. Ablation study

Human generated annotations are used as groundtruth in order to compute depth and disparity errors of different configurations. We report End-Point-Error (EPE) — the average Euclidean distance between estimated and groundtruth points and the percentage of curviline pixels with disparity error greater than 1, 3 and 5 pixels.

To investigate the effect of stereo setup on reconstruction accuracy, stereo pairs that have 2, 4, 6, 8, 10 and 12 degrees between images are used in reconstruction process and corresponding reconstruction errors are shown in Table 2. Reconstruction with small angles results in high depth errors due to heavy discretization although they have low matching error in pair. On the other hand, when the angle between images is large, matching task becomes harder and its error start dominating the total depth error. In our experiments, we found that reconstruction with images having 8 degrees in between addresses this trade-off most efficiently. Moreover, we observe an increase on the measured depth error when angular separation between images increased from 2 degrees to 4 degrees. We attribute it to heavily discretized depth values in reconstructions with extreme narrow angles as 2 degrees. In this regime, depth is encoded into a narrow pixel range on images which renders the matching task effortless since large depth changes mapped to small horizontal shifts. After passing this regime, error increase in disparity estimation can be compensated by performance gains resulted from geometry.

We trained our stereo matching network using different combinations of the loss terms introduced in the method section to tease out their respective contributions. Table 3 shows EPE and percentage of dislocations pixels with errors higher than 1, 3, 5 pixels for different loss term combinations. It may be seen that network trained with unsupervised warp loss alone performs poorly on localization of matching

Chimera software. In comparison with tomography method, which uses sequence of 45 images covering the angular range (−48°, 50°), we are using only one stereo pair with 2 degrees difference between the views.

Established tomography techniques are mainly brute-force multiview reconstruction algorithms and rely on geometric constraints to reconstruct 3D shapes. On the other hand, our stereo approach incorporates visual similarities of dislocations segments to lower the ambiguities that may occur when sufficient number of images for reliable reconstruction is not available or cumbersome to acquire.

To this end, in order to evaluate our method, we compare our stereo approach to two commonly used tomography techniques, Weighted Back Projection (WBP) and SIRT for varying number of images given as input. In Fig. 9, it is shown that our stereo reconstruction network outperforms both tomography techniques in low data margin i.e when only a few images are available. To our knowledge, our method is

Fig. 12. We show contour detection network's (UNet) output for TiAl test sample (top left) before (bottom left) and after (bottom right) adding additional low-contrast training sample (top right) to the training set. The performance significantly improved after adding only 1 image obtained from a different image sequence. It could be seen that the network's output was generally worse on low-contrast areas of the image. Adding one image having similar visual conditions to the training set eliminates detection errors on these regions and improves overall performance.

points. While network may be successfully trained with supervised disparity loss, its accuracy increases notably when combined with novel variance loss introduced in this work. We obtained best results when all three loss components combined.

4. Discussion

Deep neural networks achieved state-of-the-art results in all vision benchmarks for segmentation, object detection, depth estimation and reconstruction. Success of deep networks against its predecessors mainly lies on its capability to learn complex operations solely from available data. Therefore, the performance is dependent to quality, size and diversity of the training data. Immediate challenge in deep network training is to access large datasets. It may be seen from Fig. 12 in Appendix section that there are cases even one additional training sample could significantly increase detection accuracy of UNet on test sample. To this end, we constructed largest available annotated dislocation dataset for detection and stereo matching tasks.

As shown by our experiments, stereo reconstruction of curvilinear structures can be automatized while acquiring accuracy comparable to its established multi-view alternative, tomography. One of the advantages of our approach is that learning-based detection pipeline eliminates the need of manual thresholding of final reconstruction results unlike tomography, while it also enables us to apply structure-aware matching. It is evident that stereo matching of curvilinear structures may be learned from human labeled data by deep convolutional neural networks leveraging visual similarities between structures in stereo. Although relying on visual appearances yields accurate matches, it also limits the tilt range of operation since visual similarity of curvilinear structures rapidly degrades with increasing tilt angle. In our experiments, we have shown that a good performance may be acquired with stereo pairs having 8 degrees in between.

Higher precision reconstructions of curvilinear structures may be obtained via multiview approaches, however, it is crucial to keep required number of images in feasible limits in order to facilitate applicability. Therefore, our research will focus on multiview extention of our joint detection and matching approach in future.
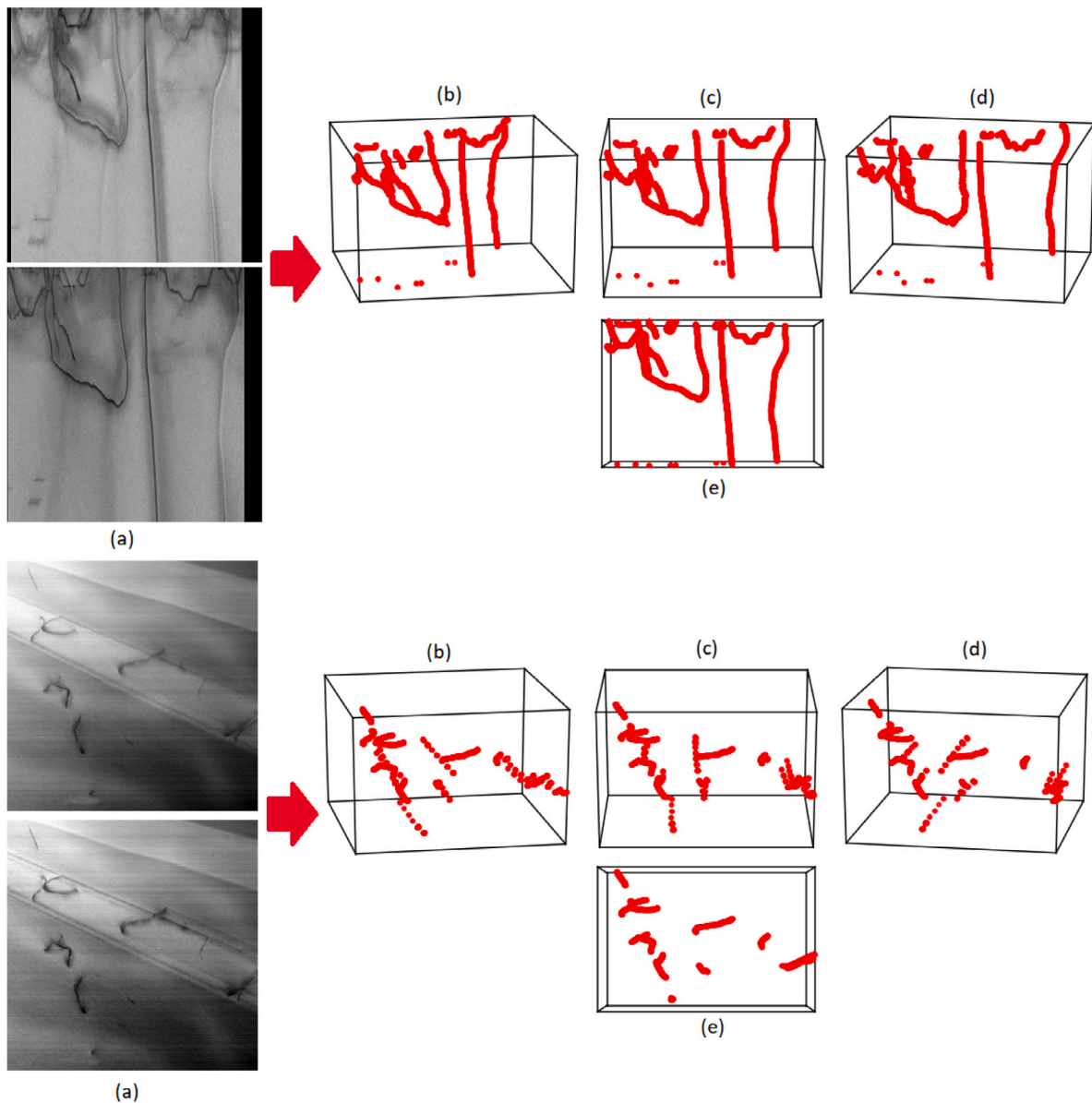
Fig. 13. Stereo pairs and 3D reconstructions of GaN (top row) and Cantor Alloy (bottom row) samples.

In conclusion, stereo-vision remains to be an acceptable alternative 3D reconstruction approach for curvilinear structures especially when acquiring large number of images is costly in terms of time and manual effort. Moreover, thanks to the recent advancements in deep neural networks, detection and matching tasks now may be combined in one neural network architecture and learned jointly. In this paper, we have therefore proposed a network architecture designed for matching curvilinear structures that delivers good performance in a wide variety of microscopy images while automatizing the process. We also introduced a novel loss term, *variance loss*, which increased our network's localization capability.

Declaration of competing interest

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

Appendix A

Neural network training

All images are resized to 512x512 and normalized to (0, 1). In order to increase dataset size, we have applied 2D image augmentations: random scaling, rotation, brightness with parameters (0.7, 1.2), (−60, 60), (0.8, 1.2) respectively. All models are optimized using the Adam

algorithm [35] with beta parameters = (0.9, 0.999) and learning rates 0.001 and 0.0001 for detection and stereo matching networks respectively. Training was performed on Tesla V100. We used the PyTorch neural network framework for our implementation.

Sample preparation and imaging conditions

- **TiAl sample.**

  Dataset of dislocations in TiAl specimen was used for training and validation, as well as to test the 3D reconstruction performed by 3D CNN.

  Dislocations were imaged in a TiAl super-alloy thin sample with nominal bulk composition of Ti-46.8Al-1.7Cr-1.8Nb (at.%). To prepare electron transparent foils from the bulk sample, the specimen was first mechanically thinned with diamond discs and electropolished (electrolyte bath composition: 5 vol.% perchloric acid, 35 vol.% 1-butanol and 60 vol.% methanol; voltage: 35 V). Then ion milling with high energy gallium ions was used to further decrease the thickness. Dislocation imaging was performed on a Thermo Scientific Tecnai Osiris transmission electron microscope operated in scanning mode (STEM) at 200 kV. Images were acquired in annular bright-field (ABF) configuration for which a 70 μm condenser aperture was used to form the probe with 12.4 mrad convergence semi-angle and detector collection angle was set to 18.4 mrad so that the rim of the direct disc (000) covers the annular detector. Tilt-series of ABF-STEM images with 1024x1024 pixels resolutions was acquired while the sample was tilted within ±50 degree tilt range and imaged every 2 degree with slightly positive deviation from the $\mathbf{g} = (002)$ two-beam diffraction condition. The sample was carefully oriented using a Fischione Instruments dual-axis tomography holder (model 2040) so that the deviation parameter $s_{\mathbf{g}}$ remains constant throughout the tilt range. To ensure the stability of the sample across all tilt angles during the tuning of the experiments, we used an alpha-wobbler with an inclination of ±10 degrees. The maximum deviation of the sample was 100 pixels, which corresponds to 300 nm. Relative precision of the sample tilt is < 0.1 degree.

- **GaN sample.**

  Dataset of dislocation images used for training and validation was acquired from a GaN sample. Details on sample preparation and imaging condition can be found in [10,36].

- **Cantor alloy sample.**

  Images of dislocations from high-entropy Cantor alloy (equimolar CrCoFeMnNi) used for training and validation. Sample has been homogenized 2 h at 1000 °C under vacuum and then 1x3 mm rectangles mechanically thinned down to 30–50 μm with SiC papers and electropolished (electrolyte bath composition: 90% ethanol and 10% perchloric acid). Dislocation imaging in Cantor alloy was performed on a Thermo Scientific Themis Titan transmission electron microscope in bright-field (BF) configuration.

Dataset specification

To be able to train models that generalize well on broad range of TEM images of dislocations, we constructed a dataset that is diverse both visually and geometrically. Our dislocation dataset consists of images of different material samples with varying imaging conditions. In Fig. 10 we present 4 data samples to show these variations in image characteristics in the dataset. Moreover, in Fig. 11 distribution of images with respect to their signal-to-noise ratio and contrast measurements are shown. We measure SnR values separately for dislocation and background regions on an image. Contrast is measured as average different in grayscale values between dislocation cores and background pixels at vicinity of dislocations. In Table 4, we show these measurements also for test samples we report our results.

Table 4

Signal-to-noise ratio and contrast values of the test samples that we used for evaluation.

Sample/Metric	SnR (disloc.) [dB]	SnR (backgr.) [dB]	Contrast
GaN	30.16	44.68	42.64
TiAl	34.24	25.32	35.78
Cantor alloy	25.70	16.42	26.91

Additional examples

In Fig. 13, we show 3D reconstructions of GaN and Cantor Alloy test samples along with the stereo pairs used for reconstruction.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ultramic.2021.113460. A cropped image sequence of the TiAl sample shown in Fig. 4 and its 3D reconstruction are provided in video format for better visualization and comparison.

References

[1] S. Amelinckx, The characterization of defects in crystals, J. Cryst. Growth (1974) 6–10.

[2] P. Hirsch, D. Cockayne, J. Spence, M. Whelan, 50 Years of TEM of dislocations: Past, present and future, Phil. Mag. 86 (29–31) (2006) 4519–4528.

[3] D.D.J. Rosier, A. Klug, Reconstruction of three dimensional structures from electron micrographs, Nature 217 (5124) (1968) 130–134.

[4] M. Weyland, P.A. Midgley, Chapter 6 electron tomography, in: Nanocharacterisation (2), The Royal Society of Chemistry, 2015, pp. 211–299.

[5] P.A. Midgley, R.E. Dunin-Borkowski, Electron tomography and holography in materials science, Nature Mater. 8 (4) (2009) 271–280.

[6] J.H. Sharp, J.S. Barnard, K. Kaneko, K. Higashida, P.A. Midgley, Dislocation tomography made easy: A reconstruction from ADF-STEM images obtained using automated image shift correction, in: Journal of Physics: Conference Series, 2008.

[7] J.S. Barnard, J. Sharp, J.R. Tong, P.A. Midgley, High-resolution three-dimensional imaging of dislocations, Science 313 (5785) (2006) 319.

[8] A. Jacome, G. Eggeler, A. Dlouhy, Advanced scanning transmission stereo electron microscopy of structural and functional engineering materials, Ultramicroscopy 122 (2012) 48–59.

[9] A. Jacome, K. Pöthkow, O. Paetsch, H.-C. Hege, Three-dimensional reconstruction and quantification of dislocation substructures from transmission electron microscopy stereo pairs, Ultramicroscopy 195 (2018) 157–170.

[10] E. Oveisi, A. Letouzey, D. Alexander, Q. Jeangros, R. Schaublin, G. Lucas, P. Fua, C. Hebert, Tilt-less 3D electron imaging and reconstruction of complex curvilinear structures, Nat. Sci. Rep. 7 (10630).

[11] E. Oveisi, A. Letouzey, S.D. Zanet, G. Lucas, M. Antoni, P. Fua, C. Hebert, Stereovision three-dimensional reconstruction of curvilinear structures imaged with a TEM, Ultramicroscopy 184 (A) (2018) 116–124.

[12] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: MICCAI, 2015.

[13] M. Hannah, Digital stereo image matching techniques, Int. Soc. Photogramm. Remote Sens. 27 (III) (1988) 280–293.

[14] J. Schönberger, H. Hardmeier, T. Sattler, M. Pollefeys, Comparative evaluation of hand-crafted and learned local features, in: Conference on Computer Vision and Pattern Recognition, 2017.

[15] G. Medioni, R. Nevatia, Segment-based stereo matching, Comput. Vis. Graph. Image Process. 31 (1) (1985) 2–18.

[16] N. Ayache, F. Lustman, Fast and reliable passive trinocular stereovision, in: International Conference on Computer Vision, 1987.

[17] A. Meygret, M. Thonnat, M. Berthod, A pyramidal stereovision algorithm based on contour chain points, in: European Conference on Computer Vision, 1990, pp. 83–88.

[18] P. Fua, A parallel stereo algorithm that produces dense depth maps and preserves image features, Mach. Vis. Appl. 6 (1) (1993) 35–49.

[19] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Pattern Anal. Mach. Intell. 23 (11) (2001) 1222–1239.

[20] G. Yang, H. Zhao, J. Shi, Z. Deng, J. Jia, Segstereo: Exploiting semantic information for disparity estimation, in: European Conference on Computer Vision, 2018, pp. 660–676.

[21] M. Ji, J. Gall, H. Zheng, Y. Liu, L. Fang, Surfacenet: An end-to-end 3D neural network for multiview stereopsis, in: International Conference on Computer Vision, 2017, pp. 2326–2334.

[22] P. Huang, K. Matzen, J. Kopf, N. Ahuja, J. Huang, Deepmvs: Learning multi-view stereopsis, in: Conference on Computer Vision and Pattern Recognition, 2018, pp. 2821–2830.

[23] A. Kar, C. Häne, J. Malik, Learning a multi-view stereo machine, in: Advances in Neural Information Processing Systems, 2017, pp. 364–375.

[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[25] M. Tan, Q.V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, 2019.

[26] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[27] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a siamese time delay neural network, in: Advances in Neural Information Processing Systems, 1993, pp. 737–744.

[28] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, End-to-end learning of geometry and context for deep stereo regression, in: International Conference on Computer Vision, 2017.

[29] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 20 (2) (2004) 91–110.

[30] R. Bolles, R. Horaud, 3DPO: A three-dimensional part orientation system, Int. J. Robot. Res. 5 (3) (1986) 3–26.

[31] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, J. Zhang, Learning for disparity estimation through feature constancy, in: Conference on Computer Vision and Pattern Recognition, 2018.

[32] J. Pang, W. Sun, J.S. Ren, C. Yang, Q. Yan, Cascade residual learning: A two-stage convolutional network for stereo matching, in: International Conference on Computer Vision Workshops, 2017.

[33] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: IJCAI, 1981, pp. 674–679.

[34] H. Hirschmüller, D. Scharstein, Evaluation of stereo matching costs on images with radiometric differences, IEEE Trans. Pattern Anal. Mach. Intell. 31.

[35] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[36] E. Oveisi, M.C. Spadaro, E. Rotunno, V. Grillo, C. Hebert, Insights into image contrast from dislocations in ADF-STEM, Ultramicroscopy 200 (2019) 139–148.