OXFORD

SuperCUT, an unsupervised multimodal image registration with deep learning for biomedical microscopy

Istvan Grexa (ID), Zsanett Zsófia Iván, Ede Migh, Ferenc Kovács, Hella A. Bolck, Xiang Zheng, Andreas Mund (ID), Nikita Moshkov,

Vivien Miczán, Krisztian Koos and Peter Horvath (ID)

Corresponding author. Peter Horvath. E-mail: horvath.peter@brc.hu

Abstract

Numerous imaging techniques are available for observing and interrogating biological samples, and several of them can be used consecutively to enable correlative analysis of different image modalities with varying resolutions and the inclusion of structural or molecular information. Achieving accurate registration of multimodal images is essential for the correlative analysis process, but it remains a challenging computer vision task with no widely accepted solution. Moreover, supervised registration methods require annotated data produced by experts, which is limited. To address this challenge, we propose a general unsupervised pipeline for multimodal image registration using deep learning. We provide a comprehensive evaluation of the proposed pipeline versus the current state-of-the-art image registration and style transfer methods on four types of biological problems utilizing different microscopy modalities. We found that style transfer of modality domains paired with fully unsupervised training leads to comparable image registration accuracy to supervised methods and, most importantly, does not require human intervention.

Keywords: unsupervised multimodal image registration; deep learning; microscopy; correlative microscopy

Istvan Grexa, Info-bionics engineer, is currently a research fellow at the Synthetic and System Biology Unit, at Biological Research Center, Szeged, Hungary. Doctoral candidate at Doctoral School of Interdisciplinary Medicine. He developed a deep learning based automatic microscopy system that helps to standardize 3D cell culture experiments. His research focuses on deep learning-based image-processing algorithms for biological applications.

Zsanett Zsófia Iván, cell and molecular biologist, research fellow at Synthetic and System Biology Unit, Biological Research Centre, Szeged, Hungary; PhD student at the Doctoral School of Biology, University of Szeged, Szeged, Hungary. Her research focuses on the cell and molecular biology, highlighting the single-cell analysis combined with light microscopy, tissue processing, AI-based image analysis and laser microdissection.

Ede Migh, earned his Master's degree in Molecular Biology from the University of Szeged and subsequently completed my Ph.D. at the Doctoral School of Biology, University of Szeged. His research focuses on spatial single-cell multi-omics, with particular interests in high-content imaging, bioimage analysis, single-cell isolation methods, and single-cell multiomics.

Ferenc Kovacs MSc, Software engineer at Single-cell technologies Ltd. with more than 14 years of experience in application development and in applied research related to image processing in the field of medical imaging and microscopy.

Hella Bolck, PhD, molecular biologist, senior research scientist at the Department of Pathology and Molecular Pathology, University Hospital Zürich, Switzerland.Her research focuses on multi-disciplinary projects that target diagnostic and prognostic challenges in human cancers. She is dedicated to expanding our understanding of the molecular basis of cancer through the integration of image analysis, machine learning, multi-omics molecular analysis and data analysis techniques thus paving the way for more precise diagnostic tools and better treatment options for cancer management.

Xiang Zheng, with over 17 years of experience in biomedical research, obtained his medical degree before immersing himself in cancer research at Peking University. He later pursued his Ph.D. at the Max Planck Institute. Currently, he is a postdoctoral researcher in Prof. Matthias Mann's group, specializing in spatial proteomics for precision medicine in cancers and neurodegenerative disorders. Proficient in advanced techniques such as multiplex immunofluorescence staining, high-resolution imaging, AI-based image analysis, single-cell laser microdissection, and ultrasensitive mass spectrometry, Xiang's research extends across human specimens, murine models, and cellular systems.

Andreas Mund, PhD, proteomics scientist decoding spatial protein expression, Associate Professor at the University of Copenhagen and CSO of OmicVision Biosciences. His vision is to pioneer a new category in precision medicine through development of Deep Visual Proteomics for single-cell spatial proteomics. This will enable better targeted therapies based on a deeper molecular understanding of diseases. His team integrates leading-edge proteomics analytics with advanced microscopy to achieve high-resolution molecular maps of protein localization and interactions within tissues. Their spatial biology approach provides an unprecedented level of certainty in identifying disease mechanisms and informing therapeutic strategies.

Nikita Moshkov received his PhD degree from the University of Szeged and Higher School of Economics (double-degree program) in 2022. He is currently a postdoc in HUN-REN Biological Research Centre, Szeged in Hungary. In 2023 he was a visiting postdoc in Broad Institute. His research interests include computer vision for biology, representation learning and bioinformatics.

Vivien Miczán, PhD, medical biotechnologist, postdoctoral researcher at the Institute of Biochemistry at HUN-REN BRC, Szeged, Hungary; She pursued her PhD in neuroscience and super-resolution imaging techniques including STORM. Her research focuses on single cell imaging and image processing to uncover molecular changes in human tumor tissue.

Krisztian Koos, PhD, computer scientist, postdoctoral researcher at BRC Szeged, Hungary. During his PhD, he developed an automated microscopy system using deep learning to study the electrophysiological properties of neurons in brain tissues. His research focuses on solving image processing problems for biological applications. Currently, he is working with vision foundation models for medical applications.

Peter Horvath, PhD, DSc, computational biologist, director of the Institute of Biochemistry at BRC Szeged, Hungary; visiting researcher at FIMM-EMBL, University of Helsinki, Finland; and group leader at the AI4Helath Institute, Helmholtz Center Munich, Germany. His group is dedicated to finding computational solutions to biological problems with an emphasis on understanding processes at a single cell level. His research focuses on the intersection of biology, engineering and computer science, and combines wet-lab and light microscopy with image analysis and machine learning methods. Our technologies are used to reveal the molecular processes of single cells and to propose personalized therapies for cancers and other human diseases.

INTRODUCTION

With the introduction of various microscopy imaging techniques, there has been an increasing demand for concurrent or consecutive image generation from the same biological sample allowing for more advanced analysis on multiple scales and integration of structural or molecular information [1, 2].

The utilization of image multiplexing has been demonstrated to be a valuable method for obtaining molecular information within whole slide tissue samples [3]. However, the acquisition of multiplexed images necessitates the repeated scanning of the entire slide, which can result in imprecise alignment of cells due to tissue reshaping caused by the application of various stains [4]. To address these issues of misalignment, automated solutions are required, as manually registering a large number of images is both time-consuming and labor-intensive.

Particular experiments might require image registration of the different imaging modalities at the cell or even subcellular levels. For instance, in the recently introduced Deep Visual Proteomics method, researchers perform single-cell laser microdissection, a procedure that requires very high precision. First, they identify target cells with high-content or tissue screening microscopy, then a deep convolutional neural network segments and selects the cells of interest [5–8]. To perform single-cell isolation, the sample is usually placed in a different microscopy setup. For such experiments, (sub)micron precision navigation between the microscopes is required [8, 9]. Due to physical reasons, such as sample drying, membrane bending on the glass slide or stage controller errors, this problem turns out to be extremely challenging without computational correction using registration algorithms [10].

Supervised state-of-the-art (SOTA) multimodal registration methods require manual image alignment for proper training, which is time-consuming; the number of expert annotators is limited, and the annotation needs to be validated. Several unsupervised image registration methods rely on pixel-wise similarity metrics like mean squared error (MSE); however, these types of networks work only on monomodal images [11]. To overcome this problem, CycleGAN [12] can be trained in an unsupervised manner to transform one modality to the other and register single modality images, but it requires a relatively large dataset [13]. Moreover, prior works demonstrate that using CycleGAN for registration in the case of microscopy images can perform poorly even if the displacement is low between the two images [14].

We propose a pipeline that uses a generative adversarial network (GAN) trained in an unsupervised manner with deep learning–based interest point detection. We compare several image registration pipelines with the SOTA methods designed for multimodal image registration, for example, Contrastive Multimodal Image Representation for Registration (CoMIR) [14, 15]. Finally, we benchmark our proposed method against the best supervised method.

We acquired four different multimodal microscopy datasets and provided landmark annotations. We also present a pipeline where the laborious manual alignment phase of the training set generation can be omitted with a reasonable trade-off in accuracy. We show that our proposed pipeline is applicable to automatic single-cell microdissection where the screening and isolation are performed with different microscopes at a micron-precise registration level.

MATERIALS AND METHODS
Datasets

- HeLa Kyoto cell culture dataset (Figure 1): The human endocervical adenocarcinoma HeLa Kyoto EGFP-alpha-tubulin/ H2B-mCherry (CLS GMBH, Cat. number.: 300670) cell line was endogenously tagged with two fluorescent proteins for the microtubule protein alpha-tubulin and the H2B histone to visualize the morphological changes in the cytoskeleton and the chromatin, respectively. In this type of dataset, our goal is to register two modalities created with different imaging techniques and conditions. Modality 1: fluorescent imaging with nuclei and cytoplasm stained in wet conditions acquired 1366×1024 RGB images with a PerkinElmer Operetta High-Content Screening system $60\times$ objective (NA 0.6). Modality 2: brightfield imaging of the same sample after drying out, acquired with a Leica LMD6 microscope $63\times$ objective (NA 0.7), resulted in 1920×1440 images. 119 image pairs from this dataset were annotated for registration, which required 16 h.

- Renal cancer tissue dataset (Figure 1): Formalin-Fixed Paraffin-Embedded (FFPE) tissue samples of human renal cell carcinomas were made available by the Department of Pathology and Molecular Pathology at the University Hospital Zürich. From selected specimens, 10 μm sections were prepared and stained with hematoxylin and eosin (H&E). Modality 1: slides were digitized with a Hamamatsu C9600 scanner equipped with $40\times$ objective 1024×1024 images. Subsequently, the coverslips were removed by immersing the samples in xylene followed by hydration using descending ethanol concentrations and air-drying. Modality 2: the same sample after the coverslip removal and drying process acquired with a Leica LMD6 microscope $63\times$ objective (NA 0.7) resulted in 1920×1440 images. 107 image pairs were annotated in this case, and the annotation process required 13 h.

- Skin tissue dataset (Figure 1): A fully anonymized skin sample with abundant lymphocytes was obtained from the Department of Pathology, Zealand University Hospital, Denmark. The sample was cut into 2.5 μm thick sections. The tissue was incubated with an anti-CD3 antibody (clone LN10, Leica Biosystems). After washing and blocking endogenous peroxidase activity, the reactions were detected using the EnVision™ Flex Magenta Chromogen system (GV900, Agilent). Finally, the slides were rinsed in water and counterstained with Mayer's hematoxylin. Modality 1: screening with a Zeiss Axioscan 7 $20\times$ objective (0.8 NA) resulted in 1200×1600 RGB images. Modality 2: imaging with LMD7 $63\times$ objective acquired 1920×1440 RGB images. Seventy-nine image pairs were annotated, which took 5 h.

- Stochastic Optical Reconstruction Microscopy (STORM) and confocal dataset (Figure 1): Contains images of mouse tissue with fluorescent immunostaining of multiple different proteins. Modality 1: confocal image of the sample obtained by a Nikon Ti-E inverted microscope equipped with a C2 scan head and a Nikon N-STORM system. Modality 2: STORM image of the same fluorophore using the same correlated setup. STORM coordinates were binned to match the resolution of the confocal image for the registration step. 126 image pairs were annotated, which took 10.5 h.

In each dataset, 80% of the images were used for training and 20% for testing.

Annotation

We constructed our ground truth set by generating image pairs from the same field of view. We marked points that can be considered the same object in both images with the MATLAB 2021b Control Point Selection toolbox [16]. This toolbox opens image pairs where the user can mark points on both images. At
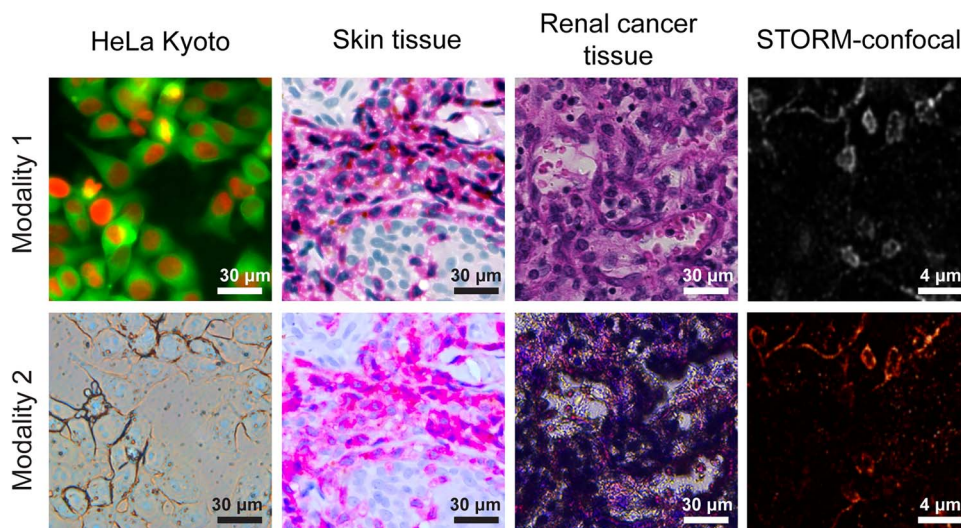
Figure 1. Representative examples from the four datasets. The image pairs are fully aligned using the annotation. In the HeLa Kyoto dataset, Modality 1 is imaged using fluorescent microscopy, and Modality 2 is an equivalent brightfield image. Skin tissue is IHC-stained and imaged in two different modalities; renal tissue is H&E-stained and imaged with different conditions. The STORM-confocal dataset is a confocal and a STORM localization microscopy image of the same fluorophore. The scale bar corresponds to 30 μm on HeLa Kyoto, skin and renal cancer tissue and 4 μm for STORM-confocal images.

least three points per image-pair were tagged as corresponding to the same area. A rigid transformation (including translation, rotation and scaling) was estimated between the images utilizing the marked points with MATLAB's 'fitgeotrans' function [17]. With this transformation, the four corner points of the image were transformed and considered as a ground truth.

Dataset preparation

For comparison of the methods, we utilized two versions of each dataset: (1) an aligned image dataset, created through the utilization of the annotations and center cropped 1024×1024 pixels from the overlapping regions, and (2) an unaligned dataset, produced by applying microscopy scaling to standardize the scale of the images, resulted in 1024×1024 images. In the case of STORM confocal, the image size was 512×512.

Style transfer methods
U-Net

The U-Net neural network is designed for biomedical image analysis. It is constructed for semantic segmentation and is capable of image-to-image translation tasks [18, 19]. Encoder–decoder style architecture of the U-Net was used as a backbone, and the final sigmoid layer was removed [20]. As a loss function, L1 distance was used, because we considered that the annotation gives a well-defined correspondence between the image pairs [21].

U-Net was trained for 1000 epochs using Adam optimizer with an initial learning rate (LR) of 3×10^{-3} and dropped by an order of magnitude every 200 epochs. A batch size of 8 was used. The images were resized to 256×256 for training and inference.

Pix2pix

Pix2pix [22] is a GAN that can be used in image-to-image translation. It consists of a generator network based on U-Net and a discriminator network that are trained in an adversarial manner: a generator is trained to generate such artificial images, so the discriminator is not able to distinguish if the images are the output of the generator or real ones. That training approach enforces the generated images to look more realistic.

Pix2pix is trained on randomly cropped out 256×256 patches from the 1024×1024 images, with a batch size of 8. For prediction, the 1024 pixel-sized images were cut into four 256×256 patches.

CycleGAN

CycleGAN [12] is a type of GAN that can be used for image-to-image translation without paired data. CycleGAN also includes a cycle consistency loss, which ensures that mapping an image from one domain to the other and back again should result in the original image.

CycleGAN models were trained for 500 epochs with LR 2×10^{-5} with single batch images. For training and inference images were resized to the size of 256×256 pixels. We could achieve visually better results with CycleGAN and CUT using a single resized image, instead of predicting and then combining separate image patches.

Contrastive unpaired translation

Contrastive unpaired translation (CUT) [23] is an improved version of CycleGAN [12], based on patch-wise contrastive and adversarial learning. This network does not require paired data from both domains to train the model. Instead, it uses a contrastive loss function that compares the representations of the translated images from the two domains to ensure that the generator produces images that are similar to those in the target domain. Mutual information of the modalities is maximized between the two images, which can result in more realistic transformed images.

CUT models were trained for 500 epochs with LR 2×10^{-5} with single batch images. For training and inferring images were resized to 256×256 pixels. Similarly to CycleGAN, transferring downscaled images produced visually more realistic microscopy images of a different modality than patch-wise style transfer.

Contrastive Multimodal Image Representation for Registration

CoMIR is a representation learning method, constructed for multimodal image registration [15]. To learn modality-independent

and rotationally equivariant representations of images, CoMIR employs a U-Net based neural network trained with a contrastive (modified InfoNCE) loss, which maximizes the mutual information between modalities. After training it transforms the different image modalities into a common latent space which reduces the registration to a monomodal task that can be registered.

CoMIR was trained with the hyperparameters proposed by the authors, but instead of 30 epochs, we used 100 epochs to ensure the convergence of the network.

For inference, both images were transformed into the CoMIR's latent space on the original image size 1024×1024. The two images in latent space were resized to 256×256 and passed to the registration algorithms.

Image registration methods

Phase cross-correlation with log-polar transform

Phase cross-correlation can be used as a method for accurately inferring the translation between two images. By utilizing the properties of the log-polar transform coordinate system and the previously obtained translation parameters, it is possible to estimate both the rotation and scaling components. The calculation of the translation parameters is performed in the Fourier space. Since the estimation of the rotation and scaling are only feasible after the moving and fixed images share a common center, first, the moving image is transformed using the previously calculated translation parameters. The images are then transformed into the log-polar space, and a phase-cross correlation is applied to obtain the rotation and scaling parameters [24].

SIFT

Scale Invariant Feature Transform (SIFT) is a robust registration method that uses scaling and rotational invariant approaches to extract key points and descriptors around the points. The algorithm uses the difference of Gaussian filters to detect key points; then, it assigns orientations based on gradient detection. These properties are calculated on both fixed and moving images and then with the RANSAC (random consensus) [25, 26] method to filter outlier points, and based on these matches, a transformation matrix can be estimated.

SuperPoint

SuperPoint is an architecture that is trained for feature and descriptor detection. The model has a single VGG16 encoder as a backbone; then, it splits into two decoder heads. The first decoder learns to detect the interest points, and the second learns the descriptors. The network is trained in a self-supervised manner; it can learn to extract the key points and descriptors using only the input images without labeling the data. Key points and descriptors are calculated on both fixed and moving images, and then, with the RANSAC method [25, 26], a transformation matrix can be estimated.

The SuperPoint model was pre-trained on Common Object in Context (COCO) [27] images and was fine-tuned in a self-supervised way. In the cases of HeLa, skin and renal cancer tissue unlabeled Modality 1 images coming from a screening microscope (see Datasets) (Figure 1) were forwarded to the MagicPoint model trained on COCO (included in the framework) for generating a pseudo ground truth. The MagicPoint model is trained through 20 000 iterations using ∼2000 pseudo ground truth images. Utilizing this MagicPoint model, new pseudo-ground truth labels were generated and repeated in the MagicPoint training process with the new labels. After training, the final labels from MagicPoint were forwarded to train Superpoint through 1 million iterations. For the STORM-confocal dataset, we used only the COCO-pretrained model as we did not have enough data to train the network.

Baselines

For baseline, we have used the previously described methods: phase cross-correlation with log-polar transform, SIFT and SuperPoint. These were used to determine if the registration was successful without using style transfer steps.

Registration pipelines

The compared registration pipelines have two steps: first, we use style transfer to transform the image pairs into the same modality. Here, images of Modality 2 are the inputs to style transfer methods and transformed into Modality 1. The only exception is CoMIR where both Modality 1 and Modality 2 images were transformed into a common latent space. All tested style transfer methods (U-Net, pix2pix, CycleGAN, CUT, CoMIR) were trained on aligned image pairs, which we consider as supervised training. Since CycleGAN and CUT are designed to work with unpaired datasets [12, 23], we trained another model using the unaligned image pair dataset, which is considered unsupervised training. In these cases, image pairs were mixed up.

Style transferred images and Modality 1 were forwarded to registration methods with 256×256 size: phase cross-correlation, SIFT and SuperPoint algorithms. These methods resulted in translation, rotation and scaling parameters (Figure 2). Translation, rotation and scaling parameters were calculated after transformation (Figure 2C).

SuperCUT

This method is a fully unsupervised method that combines SuperPoint and CUT. Training of this method requires unaligned image pairs collected from both modalities. The CUT model, which is a part of the SuperCUT method, was trained to transform images from Modality 2 on an unaligned image set to Modality 1 (see Contrastive unpaired translation section). SuperPoint network was trained on Modality 2 images. Transformed images Modality 1 were generated by applying the CUT model on Modality 2. Key points were predicted on Modality 1 and transformed into Modality 1 images and then matched with the RANSAC algorithm. Finally, an affine matrix is estimated containing translational rotational and scaling parameters.

Hardware

All neural networks were trained on a PC with an Intel Xeon E5-2620 v4 @ 2.10GHz CPU, Nvidia Titan Xp 12GB VRAM GDDR5x (founders edition, reference card), 3840 CUDA cores @1544Mhz with Pascal architecture, 32 GB DDR4 of RAM.

Metrics

The average corner error metric (Equation (1)) is the accepted metric to evaluate the accuracy of image registration [14, 15]. The transformation matrix, which is an output of every image registration method described in the paper, is applied to corner points of the input test images. The result of the metric is the average Euclidean distance between the corner points of the obtained transformation from the image registration method and the corresponding ground-truth transformation (see Annotation section).

$$\text{Average corner error} = \sum_1^4 \frac{|\, \text{Corner}_i^{\text{Ground truth}} - \text{Corner}_i^{\text{method}} \,|}{4}$$

$$(1)$$

Registration pipeline and usage in correlative microscopy

A

Single-cell isolation workflow

B

Training strategies

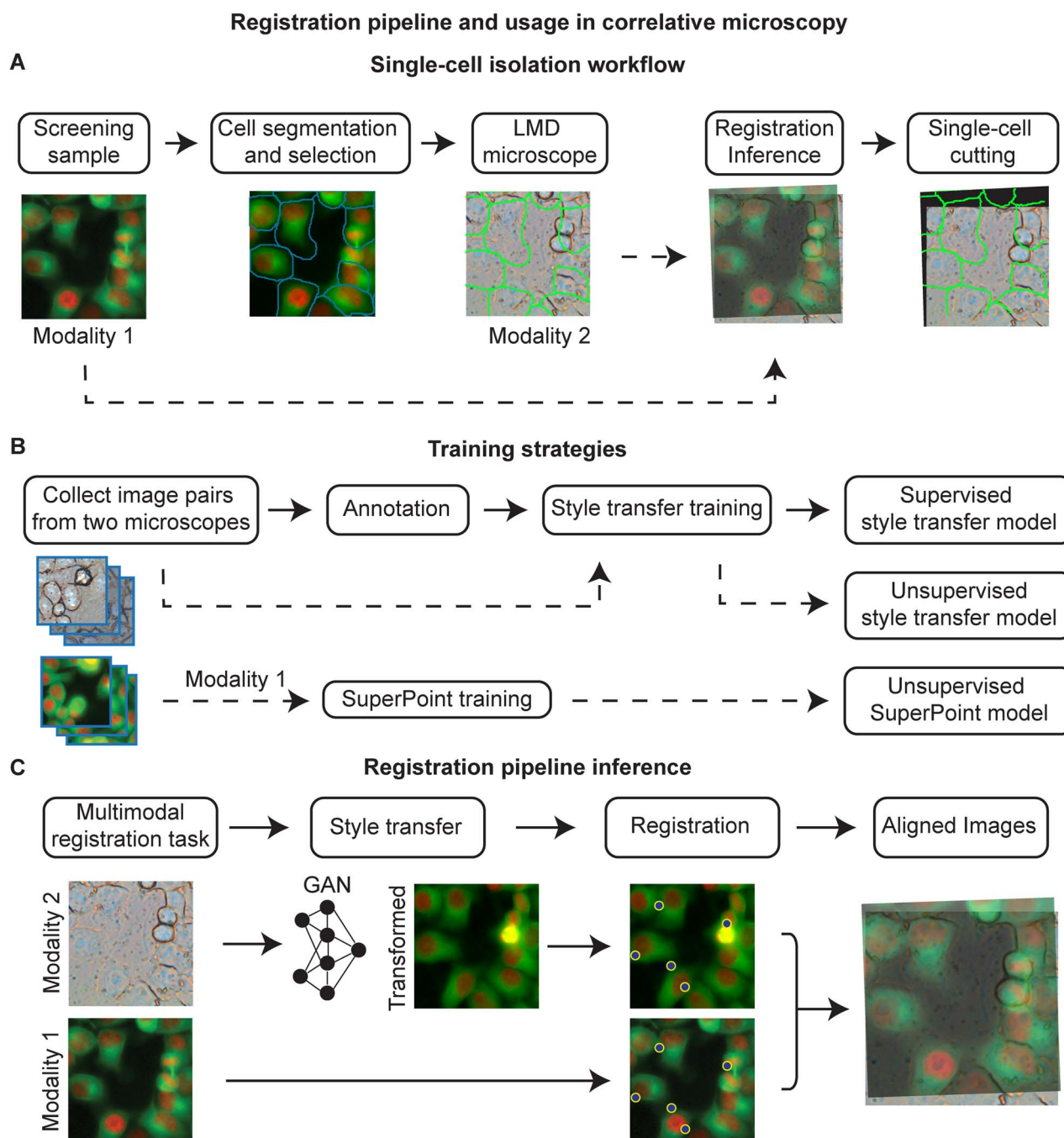**C**

Registration pipeline inference

Figure 2. (**A**) Single-cell isolation workflow. In this figure, we demonstrate the whole laser micro-dissection (LMD) process and the position of the image registration in this process. First, the images of Modality 1 images are obtained after sample screening. The images of this modality undergo single-cell segmentation to obtain the outlines of cells. The samples are then transferred to laser-cutting, where an LMD microscope is deployed (Modality 2). With image registration, the single-cell outlines obtained from the segmentation of the Modality 1 image are aligned with respect to Modality 2. After the alignment of the outlines, the laser-cutting of cells can be started. (**B**) The training strategies for style transfers and SuperPoint models are depicted. Modality 1 images from a screening microscope were passed to SuperPoint to train without labels. Supervised style transfer models were created with manually aligned images, while the unsupervised models skipped the annotation part. Unsupervised pipelines were represented with dashed lines. (**C**) Registration of a fluorescent and a brightfield image with SuperCUT. The first step is to use CUT to fabricate a fluorescent-like image (Modality 1) from the image of a laser-cutting microscope (Modality 2). SuperPoint registration calculates a rigid transformation between fabricated and genuine images.

Accepted error rates

We have set the acceptable error thresholds following our laser cutting experiments for the HeLa cell line, skin tissue and renal cancer tissue. The contour created during segmentation determines the path of the laser for microdissection. Laser ablation can cause damage along the contour depending on the settings of the laser microdissection microscope. Applying 63× magnification (NA: 0.7) objective with optimized laser power the laser causes ~4 μm damage on the sample along the contour. To protect the cells from laser-induced damage, the size of the contours was dilated with 8 μm. In these cases the accepted threshold was set to half of the enlargement: 4 μm, which is ~2.9% relative to the image size.
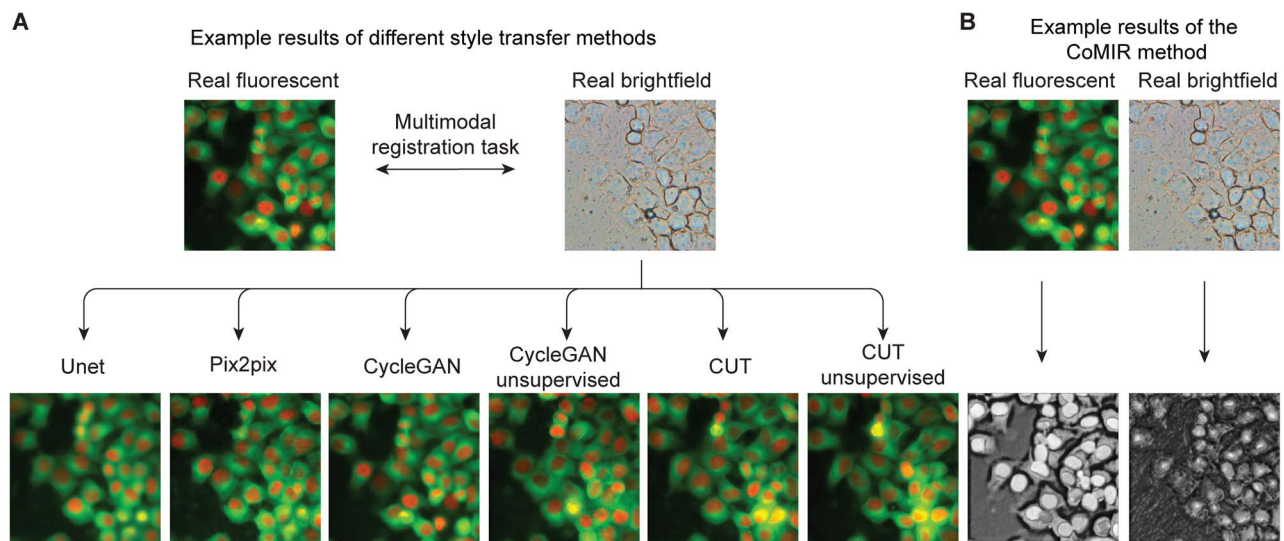
Figure 3. Illustration of several style transfer techniques. (**A**) The brightfield image transformation using the trained style transfer methods. The dense regions (CycleGAN versus unsupervised CycleGAN and CUT versus unsupervised CUT) highlight the distinction between the supervised and unsupervised approaches. (**B**) Example of a transformed image pair with CoMIR. Two images were fused into a single latent space, and then, registration techniques were used on the transformed images. Style-transferred images were passed to transformation estimation with the following registration methods: phase cross-correlation, SIFT and SuperPoint algorithms.

For the STORM-confocal dataset, we have used the currently accepted method for manual image alignment. The centers of the images were aligned also manually with dragging mode, which resulted in translation parameters only [2]. We compared the dragged image's corners with our ground truth corner points and determined the corner distance between them. The resulting errors were calculated on the test set and the average of them were used as an acceptable error of 0.6 μm, which is 1.5% relative to the image size.

Statistical analysis

Statistical analyses were performed with the Python statsmodels library [28]. The Friedman test was performed to determine if the data are different for all datasets. For *post hoc* analysis of the alignment errors, Wilcoxon signed-rank test was performed for each data pair. The significance level was set to $\alpha = 0.05$ with a 95% confidence interval. P-values were corrected for the false discovery rate using the Benjamini–Hochberg-correction method.

RESULTS

SuperCUT is designed for multimodal image registration tasks such as single-cell isolation, multiplexed imaging or intelligent high-resolution acquisition, where accurate image registration is required. We exploited the nature of these types of experiments, as high-throughput image scanning frequently results in more images from one of the modalities (Figure 2A). We conducted a comparative analysis of multiple deep learning–based image registration pipelines and our proposed method SuperCUT across four distinct datasets.

In general, the tested registration pipelines have two steps: first, style transfer was used to transform the image pairs into the same modality. Modality 2 images (Figure 1) are forwarded as inputs, and the other modality was expected as an output. Second, the output transformed image and the original Modality 1 image were registered.

We tested the following style transfer methods for modality transformation: U-Net, Pix2pix, CycleGAN, CUT and CoMIR. We used manually aligned image sets for training with all style transfer models, but in the case of CycleGAN and CUT, a model was also trained with the unaligned image sets. The results of the style transfer methods were evaluated qualitatively (Figure 3).

Comparison of supervised and unsupervised methods

The registration methods without style transfer served as baselines. We have determined an acceptable error threshold using our laser microdissection protocols for HeLa Kyoto, skin and renal cancer tissue, which is 4 μm (∼2.9%). For STORM-confocal datasets, we set the accepted error for 0.6 μm (1.5%) (see Accepted error rates) (Figure 4).

For the HeLa Kyoto dataset, the results indicate that without style transfer, the registration is unsuccessful producing very high average corner errors. The best-performing method in this case is U-Net with SuperPoint. The current state-of-the-art CoMIR method has an average performance of 11% with SIFT and 7.51% with the phase cross-correlation method (Figure 4A).

For the renal cancer tissue dataset, the lowest score was produced by U-Net with SIFT 2.03%. Errors less than the allowable error could be achieved by supervised CUT with SuperPoint, unsupervised CycleGAN with SuperPoint and unsupervised CUT with SuperPoint (SuperCUT) (Figure 4B).

For the skin tissue dataset, SIFT performance without any style transfer was the lowest error, 0.41%. U-Net with SIFT, unsupervised CycleGAN with SIFT and CoMIR with SIFT could approximate 0.41%, 0.42% and 0.43%, respectively. Errors below the allowable error range were obtained by all approaches that didn't register with phase cross-correlation (Figure 4C).

For the STORM-confocal dataset, our proposed model Super-CUT produces the lowest average error of 0.65%, among U-Net with SuperPoint, which produced 0.67% average corner error (Figure 4D).

Comparing the mean average corner errors shows that our proposed pipeline CUT with SuperPoint (SuperCUT) scores the highest among unsupervised methods and has similar performance to the best-supervised method U-Net with SuperPoint
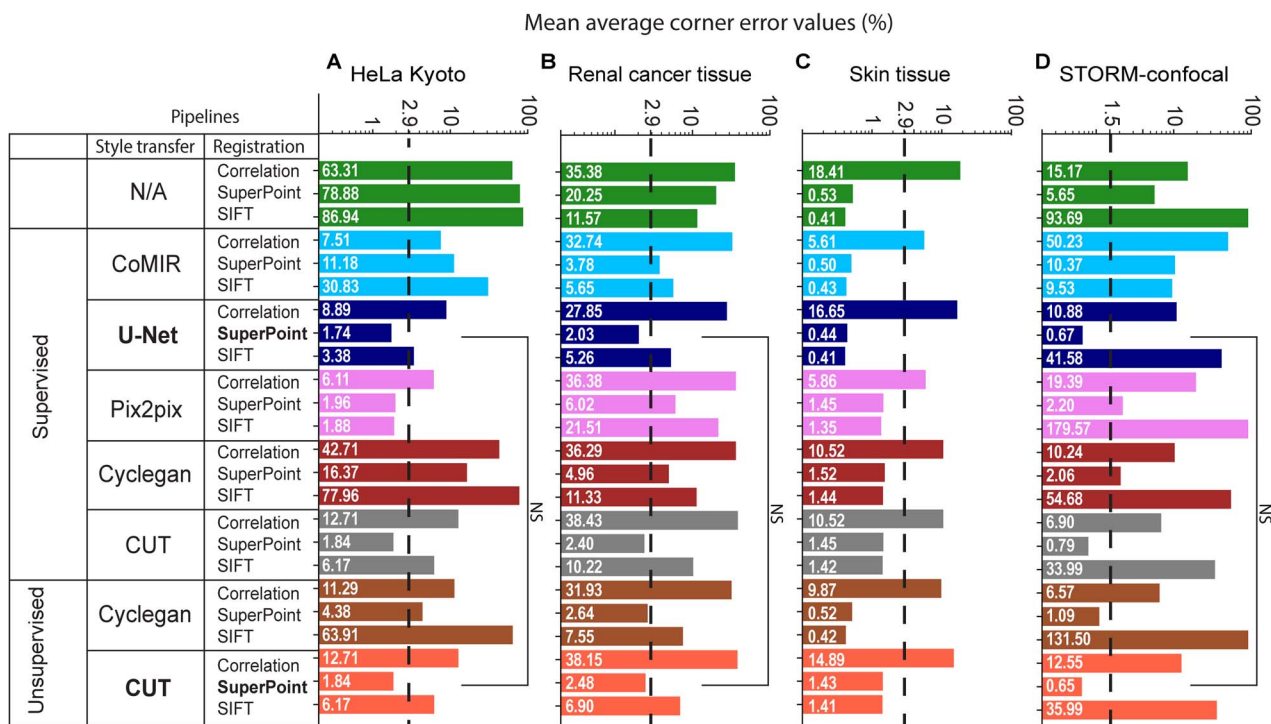
Mean average corner error values (%)

Pipelines			A HeLa Kyoto	B Renal cancer tissue	C Skin tissue	D STORM-confocal
	Style transfer	Registration				
Supervised	N/A	Correlation	63.31	35.38	18.41	15.17
		SuperPoint	78.88	20.25	0.53	5.65
		SIFT	86.94	11.57	0.41	93.69
	CoMIR	Correlation	7.51	32.74	5.61	50.23
		SuperPoint	11.18	3.78	0.50	10.37
		SIFT	30.83	5.65	0.43	9.53
	U-Net	Correlation	8.89	27.85	16.65	10.88
		SuperPoint	1.74	2.03	0.44	0.67
		SIFT	3.38	5.26	0.41	41.58
	Pix2pix	Correlation	6.11	36.38	5.86	19.39
		SuperPoint	1.96	6.02	1.45	2.20
		SIFT	1.88	21.51	1.35	179.57
	Cyclegan	Correlation	42.71	36.29	10.52	10.24
		SuperPoint	16.37	4.96	1.52	2.06
		SIFT	77.96	11.33	1.44	54.68
	CUT	Correlation	12.71	38.43	10.52	6.90
		SuperPoint	1.84	2.40	1.45	0.79
		SIFT	6.17	10.22	1.42	33.99
Unsupervised	Cyclegan	Correlation	11.29	31.93	9.87	6.57
		SuperPoint	4.38	2.64	0.52	1.09
		SIFT	63.91	7.55	0.42	131.50
	CUT	Correlation	12.71	38.15	14.89	12.55
		SuperPoint	1.84	2.48	1.43	0.65
		SIFT	6.17	6.90	1.41	35.99

Figure 4. Barplots of mean average corner error across four different datasets for each registration pipeline. Errors are measured in percentages relative to the image size. The accepted error ranges are marked with a black dashed line, respectively, to each dataset. The best supervised U-Net with SuperPoint and unsupervised pipelines [CUT with SuperPoint, (SuperCUT)] were highlighted with bold text.

(Figure 4, Supplementary Tables 1–3 available online at http://bib.oxfordjournals.org/).

Friedman's non-parametric test was used for analyzing the distribution of the samples. We found that all datasets contain significant differences. The Wilcoxon signed-rank test was performed for *post hoc* testing to find the statistical differences between the methods. Results show that in two datasets (Figure 1A and D), HeLa Kyoto and STORM-confocal, using style transfer significantly improved the results compared to SIFT and phase cross-correlation (Figure 1A and D, Supplementary Tables 1 and 3 available online at http://bib.oxfordjournals.org/). In the case of skin tissue data (Figure 1C), there is no statistical difference between SIFT and the best-performing style transfer methods: U-Net with SIFT, Unsupervised CycleGAN with SIFT and CoMIR with SIFT. We highlight that there is no significant difference between U-Net with SuperPoint and the unsupervised CUT model with SuperPoint on any dataset (Figure 4A, B and D).

Quality assessment

Based on our results on skin tissue, it is not beneficial to use registration pipelines since a classical method is capable of solving the alignment task with low errors.

We benchmarked the performance of the CUT and U-Net with a limited number of training images on three datasets: HeLa, renal tissue and STORM-confocal. We trained a CUT and U-Net model on 25-50-75 images from the available training data, respectively. Each training was repeated with five different random seeds. The results of this experiment demonstrate that image registration performance increases if more images are in the training set (Figure 5A). U-Net with SuperPoint can learn the image transformation with 50 aligned images to achieve the accepted error ranges of each dataset. SuperCUT requires the use of all of the images to have an acceptable performance.

We tested our method's tolerance for displacement on HeLa and renal tissue datasets. As we discussed, for the skin tissue dataset, it is not relevant to use the pipelines, and for the STORM-confocal dataset, we used full images to register; thus, we had to add padding to simulate displacement.

The Modality 2. images were displaced with fixed amounts of pixels and measured and calculated the mean average corner errors for each displaced dataset. With an overlap of at least 70%, the registration mean average corner error would be 2%. For the HeLa dataset, the SuperCUT was more tolerant of lower overlaps. The mean average corner error starts to increase at 60% overlap for the U-Net with SuperPoint and at 55% for SuperCUT. Below 40% overlap, both methods' performance dropped into the unacceptable range (Figure 5B). We measured the training times of the U-Net and CUT using limited amounts of images (Figure 5C). We measured the two pipelines' inference time, which showed a 0.03 s difference (Figure 5D). The inference time of each pipeline component (U-Net, CoMIR, Cyclegan, CUT, phase cross-correlation, SuperPoint) is indicated in Figure 5D.

DISCUSSION

We present a pipeline for automating the registration of images acquired from different imaging modalities, which enables precise correlative microscopy in a high-throughput manner. To train the proposed pipeline, annotated data are not required, only image pairs, typically in the order of 100 to have the best performance.

As an example, we demonstrate that the proposed method automates the process of single-cell microdissection by aligning the contours of cells, thereby reducing the need for laborious manual alignment. Automatic single-cell microdissection's greatest bottleneck is aligning the contours of the cells to their proper place on the dissection microscope manually. This is mostly due
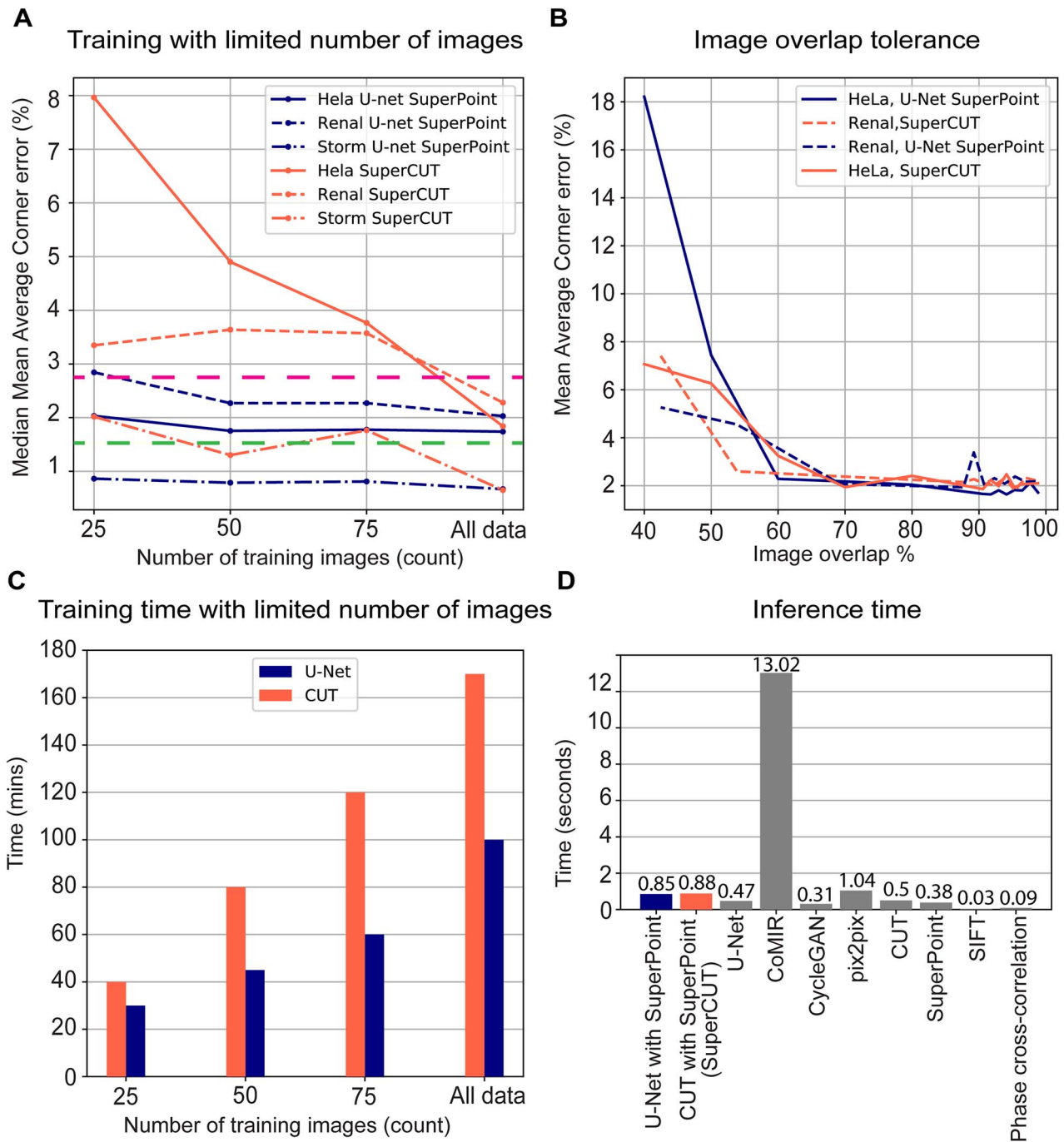
Benchmark

A Training with limited number of images

B Image overlap tolerance

C Training time with limited number of images

D Inference time

Figure 5. (**A**) Medians of mean average corner errors of runs with different seeds and for different training set sizes. U-Net (orange lines) can learn image transformation with only 50 aligned images to have errors less than the acceptable range for all datasets. SuperCUT (blue lines) needs all available training data to ensure acceptable errors. Acceptable errors were marked at ~2.9 % for HeLa and Renal and at 1.5% for STORM-confocal. (**B**) Overlap tolerance while inferencing for two datasets. (**C**) Barplots of the training time (*Y*-axis) benchmark of CUT (orange bars) and U-Net (blue bars) style transfer components of the pipelines for different sizes of the training set (*X*-axis). (**D**) Results of the inference time per registration for the best supervised and unsupervised method: U-Net with SuperPoint (blue bar) and SuperCUT (orange bar) that run on GPU. Gray bars indicated all other tested style transfer and registration methods that run on CPU. Note that CoMIR needs to make two transformations for one registration.

to the bending or drying of the sample. Using the registration algorithm, one can overcome this issue and can rapidly register a field of views containing several cells in just a few seconds. This work enables high-throughput laser microdissection with unsupervised models and may enable complete automation of training image acquisition and model training [5, 6, 29].

We also show that our pipeline can be used in other correlation microscopy techniques, such as localization microscopy, to enable automatic registration of images with high precision between STORM and confocal modalities. Our pipeline was capable of registering every test image within the acceptable error range, which enables automatic registration instead of manual image alignment [2, 30]. We believe that our method will allow researchers to rapidly generate models for the registration of other types of multiplexed images [31, 32].

Our experiments show that the precision of multimodal image registration can be significantly improved by using style transfer methods. In some cases, where the image modalities are vastly different, such as the HeLa dataset with fluorescent and brightfield images, none of the conventional registration methods worked without style transfer. By using style transfer and the SuperPoint method, we achieved significant improvements in registration compared to phase cross-correlation and SIFT in two datasets. In the case of HeLa Kyoto and STORM-confocal datasets, we observed significant improvements in registration using style transfer methods compared to no style registration methods. In the case of renal tissue, we observed slight improvements, while in the case of skin tissue, which had similar imaging modalities, we did not observe any increase in AUC values with style transfer. All other tested registration methods produced very low average corner errors, and most of them perform in the acceptable range.

Benchmarking shows that using the best supervised method U-Net with SuperPoint can have great performance even overfitting on 25 images; however, it can take up to 4 h to properly annotate those images.

We notice that it is very important to visually track how the contrastive unpaired network is learning the modality transformation because, in some cases. it is possible that the loss function is stuck and learns to generate wrong images. It might lead to very high registration errors.

Based on our limited number of experimental conditions, one of the limitations we observed is that in the case of thick tissue sections, it is difficult to train high-quality style transfer models; therefore, the registration can be imprecise (e.g. Figure 4B). Consequently, a similar effect is noticed, when the images were not in the same focal plane as those used for training the style transfer model.

Our primary finding is that it is possible to construct a registration pipeline utilizing our proposed method SuperCUT and train it without any supervision while maintaining results comparable to supervised methods.

Key Points

- We propose SuperCUT, an unsupervised pipeline that can register multimodal microscopy images
- We compare several unsupervised and supervised registration pipelines on four different microscopy datasets.
- Our unsupervised registration pipeline yields comparable results to supervised ones without any laborious annotation.

- Single-cell microdissection's bottleneck is the need for manual alignment of cellular contours, the proposed method overcomes this problem and automates single-cell isolation
- We show that the pipeline is able to register localization microscopy datasets.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

AUTHOR CONTRIBUTIONS

I.G. carried out the registration pipeline and the data analysis, V.M. and I.Z.S.Z.S. performed image annotation, E.M., X.Z., H.A.B. and F.K. made samples and images, and V.M., K.K., A.M., N.M. and P.H. conceived the project. All authors participated in writing the manuscript.

DATA AVAILABILITY

The HeLa Kyoto, Skin and Renal cancer tissue dataset with landmark annotation published at Zenodo: https://zenodo.org/record/8162985. The best SuperCUT trained models for HeLa Kyoto, Skin and Renal cancer tissue datasets were published at Zenodo: The STORM-confocal dataset is available at request.

CODE AVAILABILITY

The code for experiment reproduction and figures is available at https://github.com/grexai/unsupervised-microscopy-image-registration

ETHICS

The local ethics commission approved this study (BASEC# 201 9-01959), and all patients provided written consent.

REFERENCES

1. Haniffa M, Taylor D, Linnarsson S, *et al.* A roadmap for the human developmental cell atlas. *Nature* 2021;**597**(7875): 196–205. https://www.nature.com/articles/s41586-021-03620-1. (23 January 2023, date last accessed).

2. Barna L, Dudok B, Miczán V, *et al.* Correlated confocal and super-resolution imaging by VividSTORM. *Nat Protoc* 2016;**11**(1), pp. 163–83. https://doi.org/10.1038/nprot.2016.002. https://pubmed.ncbi.nlm.nih.gov/26716705/. (9 January 2023, date last accessed).

3. Chiaruttini N, Burri O, Haub P, *et al.* An open-source whole slide image registration workflow at cellular precision using Fiji, QuPath and Elastix. *Front Comput Sci* 2022;**3**:780026. https://doi.org/10.3389/fcomp.2021.780026.

4. Wang C-W, Ka S-M, Chen A. Robust image registration of biological microscopic images. *Sci Rep* 2014;**4**:6050.

5. Brasko C, Smith K, Molnar C, *et al.* Intelligent image-based in situ single-cell isolation. *Nat Commun* 2018;**9**(1):226.

6. Mund A, Coscia F, Kriston A, *et al.* Deep visual proteomics defines single-cell identity and heterogeneity. *Nat Biotechnol* 2022;**40**(8): 1231–40.

7. von Eggeling, von Eggeling, Melle C, Ernst G. Microdissecting the proteome. *Proteomics* 2007;**7**(16):2729–37.

8. Hollandi R, Diósdi Á, Hollandi G, *et al.* AnnotatorJ: an ImageJ plugin to ease hand annotation of cellular compartments. *Mol Biol Cell* 2020;**31**(20):2179–86.

9. Demichev V, Messner CB, Vernardis SI, *et al.* DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* 2020;**17**(1):41–4.

10. Rashidi HH, Chen M. Preface: artificial intelligence (AI), machine learning ML and digital pathology integration are the next major chapter in our diagnostic pathology and laboratory medicine arena. *Semin Diagn Pathol* 2023;**40**(2):69–70.

11. Dalca AV, Balakrishnan G, Guttag J, Sabuncu MR. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Med Image Anal* 2019;**57**:226–36.

12. Zhu J-Y, Park T, Isola P, and Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Institute of Electronics and Eletrical Engineers, 2017 IEEE International Conference on Computer Vision (ICCV)*. Italy, 2017, pp. 2242–51. https://doi.org/10.1109/iccv.2017.244.

13. Arar M, Ginger Y, Danon D, *et al.* Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA, 2020, pp. 13407–16. https://doi.org/10.1109/cvpr42600.2020.01342.

14. Lu J, Öfverstedt J, Lindblad J, Sladoje N. Is image-to-image translation the panacea for multimodal image registration? A comparative study. *PloS One* 2022;**17**(11):e0276196.

15. Pielawski N, Wetzer E, Öfverstedt J, *et al.* "CoMIR: contrastive multimodal image representation for registration," In: *Advances in Neural Information Processing Systems* 2020;**33**:18433–44. https://proceedings.neurips.cc/paper_files/paper/2020/file/d6428eecbe0f7dff83fc607c5044b2b9-Paper.pdf.

16. The MathWorks Inc. *Statistics and Machine Learning Toolbox Documentation*. Natick, Massachusetts: The MathWorks Inc. "cpselect." 2023. https://www.mathworks.com/help/images/ref/cpselect.html. (7 June 2023, date last accessed)

17. The MathWorks Inc. *Statistics and Machine Learning Toolbox Documentation*. Natick, Massachusetts: The MathWorks Inc. "fitgeotrans." 2023. https://www.mathworks.com/help/images/ref/fitgeotrans.html. (7 June 2023, date last accessed)

18. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*, vol **9351**. Springer, Cham. pp. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28.

19. Osman AFI, Tamam NM. Deep learning-based convolutional neural network for intramodality brain MRI synthesis. *J Appl Clin Med Phys* 2022;**23**(4):e13530. https://doi.org/10.1002/acm2.13530. https://pubmed.ncbi.nlm.nih.gov/35044073/ (15 July 2023, date last accessed).

20. Wieslander H, Gupta A, Bergman E, *et al.* Learning to see colours: biologically relevant virtual staining for adipocyte cell images. *PloS One* 2021;**16**(10):e0258546. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0258546&type=printable. (7 November 2023, date last accessed).

21. Zhao H, Gallo O, Frosio I, Kautz J. Loss functions for image restoration with neural networks. *IEEE Trans Comput Imaging* 2017;**3**(1):47–57. http://ieeexplore.ieee.org/document/7797130/.

22. Isola P, Zhu J-Y, Zhou T, and Efros AA. Image-to-image translation with conditional adversarial networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA, 2017, pp. 5967–76. https://doi.org/10.1109/cvpr.2017.632.

23. Park T, Efros AA, Zhang R, and Zhu J-Y. Contrastive learning for unpaired image-to-image translation. In: Vedaldi A, Bischof H, Brox T, Frahm JM (eds) *Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science*, vol **12354**. Springer, Cham. pp. 319–45. https://doi.org/10.1007/978-3-030-58545-7_19.

24. Reddy BS, Chatterji BN. An FFT-based technique for translation, rotation, and scale-invariant image registration. *in IEEE Transactions on Image Processing* 1996;**5**(8):1266–71. https://doi.org/10.1109/83.506761.

25. Lowe DG. Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Kerkyra, Greece, 1999, vol. **2**, pp. 1150–57. https://doi.org/10.1109/iccv.1999.790410.

26. Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Read Comput Vis* 1987;726–40. https://doi.org/10.1016/b978-0-08-051581-6.50070-2.

27. Lin T-Y, Maire M, Belongie S, *et al.* Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) *Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*, vol **8693**. Springer, Cham. pp. 740–55. https://doi.org/10.1007/978-3-319-10602-1_48.

28. Seabold, Skipper, and Josef Perktold. "statsmodels: Econometric and statistical modeling with python." *Proceedings of the 9th Python in Science Conference*. 2010. https://www.statsmodels.org/stable/generated/statsmodels.distributions.empirical_distribution.ECDF.html#statsmodels.distributions.empirical_distribution.ECDF. (9 June 2023, date last accessed).

29. Mund A, Brunner A-D, Mann M. Unbiased spatial proteomics with single-cell resolution in tissues. *Mol Cell* 2022;**82**(12): 2335–49.

30. Tam J, Cordier GA, Bálint Š, *et al.* A microfluidic platform for correlative live-cell and super-resolution microscopy. *PloS One* 2014; **9**(12):e115512. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0115512&type=printable (10 November 2023, date last accessed).

31. Windhager J, Zanotelli VRT, Schulz D, *et al.* An end-to-end workflow for multiplexed image processing and analysis. *Nat Protoc* 2023;**18**(11):3565–613. https://www.nature.com/articles/s41596-023-00881-0. (10 November 2023, date last accessed).

32. Gatenbee CD, Baker AM, Prabhakaran S, *et al.* Virtual alignment of pathology image series for multi-gigapixel whole slide images. *Nat Commun* 2023;**14**(1):4502–14. https://www.nature.com/articles/s41467-023-40218-9. (10 November 2023, date last accessed).