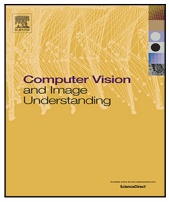
Computer Vision and Image Understanding 240 (2024) 103939

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Unsupervised deep learning of foreground objects from low-rank and sparse dataset

Keita Takeda *, Tomoya Sakai

School of Information and Data Sciences, Nagasaki University, 1-14 Bunkyo, Nagasaki City 852-8521, Japan

ARTICLE INFO

MSC:
62H35
68T45
68U10

Keywords:
Nuclear loss
Dual frame U-Net
Low-rank and sparse model
Background subtraction

ABSTRACT

Foreground object identification can be considered as anomaly detection in a redundant background. This paper proposes unsupervised deep learning of foreground objects on the basis of the prior knowledge about spatio-temporal sparseness and low-rankness of foreground objects and background scenes. The proposed framework trains a U-Net model to encode and decode the sparse foreground objects in batches of input images with low-rank backgrounds, by minimizing a combination of nuclear and ℓ_1 norms as a loss function. This approach is similar to background subtraction based on robust principal component analysis (RPCA): an iterative method that detects sparse foreground objects as outliers while learning the principal components of the linearly dependent background. In contrast, the proposed method is advantageous over RPCA in that once the U-Net model has learned enough features common to the foreground objects, it can robustly detect them from any single image regardless of the low-rankness and sparseness. The U-Net also enables online object segmentation with much less computational expense than that of RPCA. These advantages are illustrated with background subtraction in video surveillance. It is also shown that the proposed method can build up a well-generalized cell segmentation model from only a few dozen unannotated training images.

1. Introduction

Our objective is to obtain informative image features with minimal supervision, by leveraging prior knowledge. It is observed that structural events or spatially structured patterns in images often result in a sequence or set of images that exhibit recurring and redundant features. On the other hand, unusual or abnormal events of interest tend to cause sparse outlying features. In many cases, the combination of such redundant and sparse features can be represented by a low-rank and sparse (L+S) model.

The L+S modeling has proven to be a successful approach for analyzing various types of time-series data. This includes surveillance videos that capture both the background and foreground elements (Candès et al., 2011; Guyon et al., 2012), optical-flow sequences that involve egomotion and object motion (Sakai and Kuhara, 2015), ultrasound images with clutter and blood flow (Zhang et al., 2020), music with accompaniment and singing voice components (Huang et al., 2012), and even respiratory auscultation sounds (Sakai et al., 2016). These instances underscore the inherent versatility of the L+S model, positioning it as an invaluable asset for processing and analyzing a collection of high-dimensional data across diverse domains.

The paradigmatic use of the L+S model manifests in robust principal component analysis (RPCA) (Skočaj et al., 2007; Candès et al., 2011), a powerful technique designed to effectively capture the low-rank structure of a high-dimensional dataset by isolating sparse outliers. This analytical framework has been further fortified by the development of efficient algorithms tailored for RPCA or L+S approximation (Bouwmans and Zahzah, 2014). Beyond its classical applications, the L+S modeling finds resonance in the realm of deep learning techniques. G-LBM (Rezaei et al., 2020) employs a variational auto-encoder to learn the background as a low-dimensional manifold and segment foreground objects as outliers. Another innovative approach, CORONA (Solomon et al., 2020), inserts trainable convolutional layers into the computation graph of the iterative shrinkage/thresholding algorithm applied to RPCA for ultrasound image processing. A recent proposal (Cai et al., 2021) presents an efficient, learnable RPCA variant, circumventing the need for singular value thresholding in the computation graph.

While all these existing methodologies based on the L+S model can separate sparse foreground from a low-rank background without requiring annotations, they share a common limitation — they exclusively focus on learning background components, neglecting the distinctive features of foreground components. This drawback becomes evident in scenarios where the focus should be on extracting features specific to foreground components. As an illustrative example, consider the context of surveillance videos targeting foreground pedestrians; background components learned in one scene may not be suitable for background subtraction in videos taken from different viewpoints, even

* Corresponding author.
E-mail address: ktakeda@nagasaki-u.ac.jp (K. Takeda).

https://doi.org/10.1016/j.cviu.2024.103939
Received 7 May 2023; Received in revised form 15 December 2023; Accepted 12 January 2024
Available online 18 January 2024

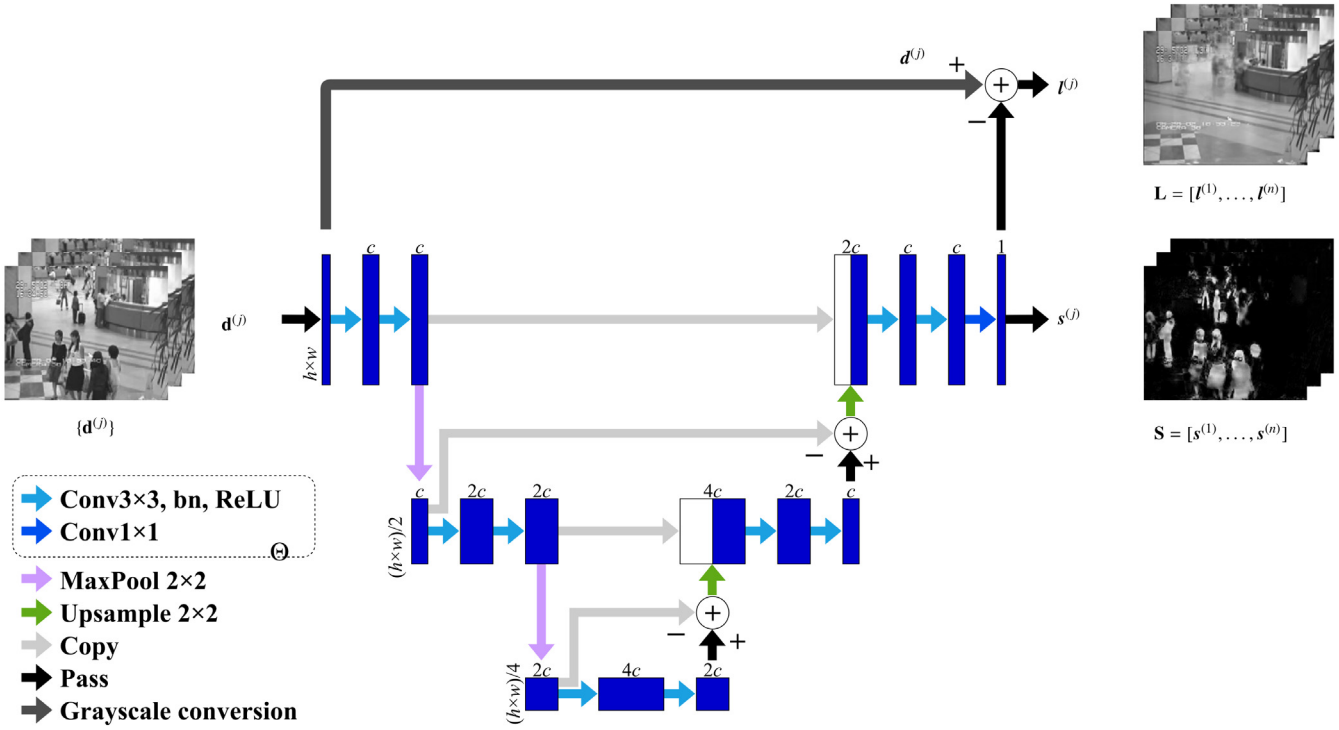


Fig. 1. Our U-Net-based model for online foreground separation. The model takes an input image composed of m pixel values represented by $\boldsymbol{d}^{(j)} \in \mathbb{R}^m$. Using the dual-frame U-Net architecture (Han and Ye, 2018), the model estimates the sparse foreground image as $\boldsymbol{s}^{(j)} \in \mathbb{R}^m$. Each box with a number in the figure corresponds to a tensor of feature maps and its number of channels. The matrices $\mathbf{L}, \mathbf{S} \in \mathbb{R}^{m \times n}$ store the output images as column vectors $\boldsymbol{l}^{(j)} = \boldsymbol{d}^{(j)} - \boldsymbol{s}^{(j)}$ and $\boldsymbol{s}^{(j)}$ for $j = 1, \ldots, n$. During the training phase, the model's parameters Θ are optimized to minimize the sum of nuclear and ℓ_1 norms, as shown in Eq. (9). This minimization results in $\mathbf{L}$ and $\mathbf{S}$ being respectively as low-rank and sparse as possible.
Source: Extracted from our DICTA paper (Takeda et al., 2022)

when the foreground comprises consistent objects like pedestrians. In the domain of foreground object segmentation, where supervised deep learning stands as the mainstream approach (Kalsotra and Arora, 2022), there exists untapped research potential for exploring the incorporation of low-rank and sparse priors. Such exploration could lead to the development of methodologies for unsupervised learning, specifically tailored to extract features pertinent to the sparse foreground of interest in situations where annotated training datasets are scarce or impractical.

We introduce an innovative approach for semantic object segmentation using unsupervised deep learning that capitalizes on low-rank and sparse priors for training. Our method leverages a popular hourglass-like model with convolutional layers, illustrated in Fig. 1, which is adept at encoding and decoding local and sparse foreground features. To train our model, we adopt the sum of nuclear and ℓ_1 norms, which serves as the objective function of RPCA, as the loss function to minimize with respect to the model parameters. In comparison to the other recent deep learning-based methods mentioned above, our method stands out as the most succinct and pragmatic approach for identifying learned sparse foregrounds amidst cluttered backgrounds. Our model can learn the features of sparse components of training datasets without supervision, making it ideal for unsupervised segmentation tasks. After acquiring sufficient knowledge about the features of foreground objects, our model can isolate them in any given scene, irrespective of their level of sparseness. Hence, there is no need to retrain the model to suit different backgrounds, resulting in computationally efficient online segmentation of target objects whose features have been assimilated as the sparse components of training datasets. We show these advantages in background subtraction and foreground object segmentation tasks. We also prove the generalization capability of our model through cell segmentation, where it was trained with a small-scale dataset.

Key contributions of this article include:

- development of an unsupervised deep learning method for online detection of foreground objects,
- introduction of a novel loss function combining nuclear and sparsity-inducing norm,
- outperformance over the traditional robust principal component analysis (RPCA) in both computational efficiency and output quality, and
- demonstration of high generalizability utilizing limited amounts of unannotated data in the application to video surveillance and cell segmentation.

This article serves as an extension to our conference paper presented at DICTA 2022 (Takeda et al., 2022). Novel contributions of this article include:

- comparison on sparsity-inducing loss functions ℓ_1 and SCAD,
- additional evidence that our model learns discriminative features of foreground objects without supervision, and
- intensive evaluation on the segmentation generalizability regarding dataset size and diversity.

2. Low-rank and sparse (L+S) model

Consider a batch of n m-dimensional vector data denoted by $\{\boldsymbol{d}^{(j)}\}$ where each vector $\boldsymbol{d}^{(j)} \in \mathbb{R}^m$. Assume that $\{\boldsymbol{d}^{(j)}\}$ can be decomposed into two distinct batches: $\{\boldsymbol{l}^{(j)}\}$ and $\{\boldsymbol{s}^{(j)}\}$. The first batch $\{\boldsymbol{l}^{(j)}\}$ has redundancy, meaning that the vectors in this batch are linearly dependent on each other. We can measure this dependence using an m by n matrix $\mathbf{L} = [\boldsymbol{l}^{(1)}, \ldots, \boldsymbol{l}^{(n)}] \in \mathbb{R}^{m \times n}$, which has low-rankness. On the other hand, the second batch $\{\boldsymbol{s}^{(j)}\}$ has a sparse nature, meaning that most of its entries are zero. We can measure the sparsity of this batch by counting the number of nonzero entries in the m by n matrix $\mathbf{S} = [\boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(n)}] \in \mathbb{R}^{m \times n}$.

The L + S model is a technique used to represent a given matrix $\mathbf{D} = [\boldsymbol{d}^{(1)}, \ldots, \boldsymbol{d}^{(n)}] \in \mathbb{R}^{m\times n}$ as the sum of a low-rank matrix $\mathbf{L}$ and a sparse matrix $\mathbf{S}$. Finding the best L + S model for a given $\mathbf{D}$ is a constrained multi-objective optimization problem. The objective is to minimize both the rank of $\mathbf{L}$ and the ℓ_0 norm of $\mathbf{S}$, subject to the constraint that $\mathbf{D} = \mathbf{L} + \mathbf{S}$:

$$\underset{(\mathbf{L},\mathbf{S})}{\text{Minimize}} \left\{\text{rank}\,\mathbf{L}, \|\mathbf{S}\|_0\right\} \quad \text{subject to} \quad \mathbf{D} = \mathbf{L} + \mathbf{S}. \tag{1}$$

The rank of $\mathbf{L}$, rank$\mathbf{L}$, represents the maximal number of linearly independent columns of $\mathbf{L}$, while the ℓ_0 norm of $\mathbf{S}$, $\|\mathbf{S}\|_0$, counts the number of nonzero entries in $\mathbf{S}$. The goal is to find the pair of matrices that best approximates $\mathbf{D}$ by recovering a low-rank matrix from the data corrupted by sparse, unknown errors. This approach is known as robust principal component analysis (RPCA). Unlike classical PCA, which assumes small and dense noise, the entries in $\mathbf{S}$ can have arbitrarily large magnitudes, and its support is assumed to be sparse but unknown beforehand. Unfortunately, the optimization problem in Eq. (1) is computationally intractable.

Fortunately, it has been shown that the optimal $(\mathbf{L}, \mathbf{S})$ for the problem described in Eq. (1) can be obtained by solving the following convex optimization problem (Candès et al., 2011):

$$\underset{(\mathbf{L},\mathbf{S})}{\text{Minimize}} \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 \quad \text{subject to} \quad \mathbf{D} = \mathbf{L} + \mathbf{S}. \tag{2}$$

Here, the nuclear norm of $\mathbf{L}$, $\|\mathbf{L}\|_*$, is defined as the sum of the singular values of $\mathbf{L}$, whereas the matrix ℓ_1 norm of $\mathbf{S}$, $\|\mathbf{S}\|_1$, is defined as the sum of the absolute values of the matrix entries. Minimizing the nuclear norm $\|\mathbf{L}\|_*$ and the matrix ℓ_1 norm $\|\mathbf{S}\|_1$ encourages low-rankness and sparsity in $\mathbf{L}$ and $\mathbf{S}$, respectively, as these norms can be considered as convex envelopes of the matrix rank and ℓ_0 norm. The hyperparameter $\lambda > 0$ balances the contributions of these norms to the minimizers.

The objective function in Eq. (2) is non-differentiable at the low-rank and sparse matrices $(\mathbf{L}, \mathbf{S}) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n}$, as it involves the non-differentiable nuclear norm and ℓ_1 norm. Therefore, gradient-based methods are unable to provide optimal low-rank and sparse matrices, but can only obtain their coarse approximations. To tackle this issue, practical algorithms for RPCA via principal component pursuit (Bouwmans and Zahzah, 2014) are typically designed based on proximal methods. For instance, one can use the alternating directions method of multipliers (ADMM) (Gabay and Mercier, 1976; Boyd et al., 2011) to derive the following iteration steps:

$$\mathbf{L} \leftarrow \text{svt}(\mathbf{D} - \mathbf{S} - \mathbf{E}, 1/\rho), \tag{3}$$

$$\mathbf{S} \leftarrow \text{soft}(\mathbf{D} - \mathbf{L} - \mathbf{E}, \lambda/\rho), \text{and} \tag{4}$$

$$\mathbf{E} \leftarrow \mathbf{E} + \mathbf{L} + \mathbf{S} - \mathbf{D}. \tag{5}$$

Here, ρ is an arbitrary positive constant. In Eq. (3), the singular value thresholding operation (Cai et al., 2010; Ma et al., 2011), denoted as svt, is defined as

$$\text{svt}(\mathbf{Q}, \tau) = \mathbf{U}\,\text{soft}(\mathbf{K}, \tau)\mathbf{V}^\top, \tag{6}$$

where $\mathbf{U}$, $\mathbf{K}$, and $\mathbf{V}$ are the singular value decomposition (SVD) of $\mathbf{Q} = \mathbf{U}\mathbf{K}\mathbf{V}^\top$, and $\tau \geq 0$ is a threshold applied to the singular values which make up the diagonal matrix $\mathbf{K}$. The soft thresholding operation (Daubechies et al., 2004; Donoho, 1995), denoted as soft in Eqs. (4) and (6), is defined as

$$\text{soft}(q, \tau) = \text{sign}(q)\max(|q| - \tau, 0), \tag{7}$$

and works element-wise on matrices. The thresholding operations, svt and soft, correspond to the proximal mappings with respect to the nuclear and ℓ_1 norms, respectively.

The ADMM algorithm is a powerful optimization technique that can provide solutions with guaranteed convergence rates and is widely used for RPCA problems. However, the RPCA method involves computationally intensive SVD in the iteration step, even though the ADMM algorithm can produce an acceptable approximate solution with only a few hundred iterations in practice. Furthermore, RPCA does not provide any insight into the features of the sparse component.

An alternative approach to estimate the sparse component $\boldsymbol{s}$ of a test input $\boldsymbol{d}$ is to remove the principal components represented by the columns of the matrix $\mathbf{U}$. This can be done by decomposing $\boldsymbol{d}$ into its vector projection $\boldsymbol{d}_\parallel = \mathbf{U}\mathbf{U}^\top\boldsymbol{d}$ and vector rejection $\boldsymbol{d}_\perp = \boldsymbol{d} - \boldsymbol{d}_\parallel$ with respect to the principal subspace of $\{\boldsymbol{l}^{(j)}\}$. Assuming that $\boldsymbol{d} = \boldsymbol{l} + \boldsymbol{s}$, where the principal components can express $\boldsymbol{l}$ as $\mathbf{U}\boldsymbol{a}$, we can estimate $\boldsymbol{s}$ as the vector rejection $\boldsymbol{d}_\perp$ since

$$\boldsymbol{d}_\perp = (\boldsymbol{l} + \boldsymbol{s}) - \mathbf{U}\mathbf{U}^\top(\mathbf{U}\boldsymbol{a} + \boldsymbol{s}) = \boldsymbol{s} - \mathbf{U}\mathbf{U}^\top\boldsymbol{s}. \tag{8}$$

If $\boldsymbol{s}$ is sparse, and therefore the vector projection of $\boldsymbol{s}$ onto the principal subspace, given by $\mathbf{U}\mathbf{U}^\top\boldsymbol{s}$, is negligibly small, the vector projection and rejection, $\boldsymbol{d}_\parallel$ and $\boldsymbol{d}_\perp$, respectively approximate $\boldsymbol{l}$ and $\boldsymbol{s}$ well.

3. Unsupervised deep learning with low-rank and sparse priors

3.1. U-net-based model

We propose a novel deep neural network model for detecting foreground objects in images. The architecture of our model is illustrated in Fig. 1. Given a batch of input images $\{\mathbf{d}^{(j)}\}$, where each $\mathbf{d}^{(j)} \in \mathbb{R}^{m\times c_d}$ is a column vector storing m pixel values of the jth image with c_d color channels, our model takes $\mathbf{d}^{(j)}$ as input, processes it as a two-dimensional image, and outputs the corresponding foreground and background images as $\boldsymbol{s}^{(j)}$ and $\boldsymbol{l}^{(j)}$ in grayscale, respectively. The foreground objects are detected as the non-zero entries of $\boldsymbol{s}^{(j)}$.

There are a variety of hourglass-like model (Siddique et al., 2021) to encode and decode convolutional image features to produce $\boldsymbol{s}^{(j)}$. We employ a dual-frame U-Net (Han and Ye, 2018) with a set of learnable parameters, Θ. The background image $\boldsymbol{l}^{(j)}$ is obtained by subtracting the foreground image $\boldsymbol{s}^{(j)}$ from the grayscale-converted input image $\boldsymbol{d}^{(j)}$. The dual-frame U-Net is a convolutional neural network with an hourglass structure. One of its unique features is that it subtracts the features right before upsampling in the decoder from the features right after pooling in the encoder at each level. The encoder in U-Net loses the fine structure of an input image with each pooling operation while processing the remaining coarse structure for feature extraction. The dual-frame U-Net addresses this imbalance between the coarse and fine structure, making it appropriate for learning foreground object with fine structure.

It is important to note that the optimal number of channels and layers for our model may vary depending on the target application. Unlike RPCA, our model can be designed to handle multiple channels of image input, including color, providing greater flexibility and versatility for a wider range of applications.

3.2. Training by inducing low-rankness and sparseness

We employ the U-Net-based model, shown in Fig. 1, to perform foreground-background separation. Given a batch of training images, denoted as $\{\mathbf{d}^{(j)}\}$ with $j = 1, \ldots, n$, we construct matrices $\mathbf{L}, \mathbf{S} \in \mathbb{R}^{m\times n}$ by using the corresponding batches of model outputs, $\{\boldsymbol{l}^{(j)}\}$ and $\{\boldsymbol{s}^{(j)}\}$, respectively, as their columns. Following the RPCA approach, we optimize the model parameters Θ by minimizing the objective function in Eq. (2) as the loss function:

$$\underset{\Theta}{\text{Minimize}} \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1, \tag{9}$$

where $\mathbf{D} = \mathbf{L} + \mathbf{S}$ is guaranteed by the model's architecture. The minimization of nuclear norm and ℓ_1 norm encourages low-rankness and sparseness of $\mathbf{L}$ and $\mathbf{S}$, respectively.

There are other proposed functions that promote sparsity. One such function is the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), which connects the ℓ_1 and ℓ_0 norms as

$$g(x) = \begin{cases} \lambda|x| & (|x| \leq \lambda) \\ \frac{x^2 - 2a\lambda|x| + \lambda^2}{2(a-1)} & (\lambda < |x| \leq a\lambda) \\ \frac{1}{2}(a+1)\lambda^2 & (a\lambda < |x|) \end{cases} \tag{10}$$

where $a \approx 3.7$ is recommended. The sparsity-inducing function $g(x)$ serves as a substitute for the ℓ_1 norm in the optimization of model parameters Θ:

$$\underset{\Theta}{\text{Minimize}} \|\mathbf{L}\|_* + \sum_{i=1}^{m}\sum_{j=1}^{n} g(S_{ij}). \tag{11}$$

Here, S_{ij} is the ij-th entry of $\mathbf{S}$. In contrast to the ℓ_1 norm, which employs the penalty function $g(x) = \lambda|x|$, the SCAD function quantifies the number of non-zero matrix entries rather than their magnitudes: for any non-zero x value exceeding $a\lambda$, the SCAD function $g(x)$ remains constant disregarding the magnitude of x.

It is worth noting that the proposed approach is unsupervised learning, which does not require ground truth output for training. Although it is theoretically reasonable to set $\lambda = 1/\sqrt{\max(m,n)}$ (Candès et al., 2011) in Eq. (2), we suggest performing cross-validation with ground truth evaluation to find the better value for λ in Eqs. (9) and (11), and other hyperparameters such as the number of channels in the U-Net.

For the sake of convenience, we have implemented the U-Net-based model as shown in Fig. 1, and optimize its parameters using an autodiff system, PyTorch (Paszke et al., 2019) with the loss function defined in Eq. (9). To introduce the nuclear norm, we use the SVD computation which is provided as an automatically differentiable function by the autodiff system. However, the SVD computation is not required during the inference stage after the model has been trained. Instead, the model simply computes the foreground image $\boldsymbol{s}$ through the forward propagation of the dual-frame U-Net from any input image $\mathbf{d}$. The input does not necessarily have to consist of multiple images, possess low-rank or sparse characteristics, or be restricted to grayscale. Our model is capable of producing $\boldsymbol{s}$ on the basis of the learned features. Since the nonzero pixels of the output $\boldsymbol{s}$ from the dual-frame U-Net indicate the regions of foreground objects, the foreground segmentation is achieved by simply binarizing the absolute values of $\boldsymbol{s}$.

4. Experimental evaluation

4.1. Comparison with RPCA in background subtraction

We evaluated our unsupervised deep learning approach for online background subtraction by comparing it with RPCA. We conducted the evaluation on the "airport" sequence (Li et al., 2004), which consists of 3,584 color images of size 144 × 176, numbered from #1,000 to #4,583, depicting surveillance footage of pedestrians as foreground objects. For the fair comparison to RPCA, we converted the colors to grayscale values ranging from 0 to 1 and randomly select 50 frames from #1,000 to #1,600 for training, ensuring that none of them include pedestrians that appear in the remaining frames for validation and test. To evaluate the performance, we used 20 images with ground truth annotation of the foreground pedestrian regions. Ten of these images were used to validate hyperparameters, while the remaining ten were used to test the foreground segmentation performance. Another experiment with colored images is shown at later Section 4.3.

Our U-Net-based model has the same architecture as shown in Fig. 1, and treats the training images as a tensor of size 50×1×144×176. To compute the loss function, the output tensors are organized into matrices $\mathbf{L}$ and $\mathbf{S}$ with $m = 144 \times 176 = 25{,}344$ dimensions and $n = 50$ column vectors of pixel values of corresponding output images. We used Adam optimizer (Kingma and Ba, 2015) to obtain the optimal model parameters Θ for the dual-frame U-Net. We trained two U-Net models to minimize the loss function as shown in Eqs. (9) and (11). We refer to these models as "U-Net-ℓ_1" and "U-Net-SCAD", respectively.

We employed Optuna (Akiba et al., 2019) to adjust the hyperparameters: the number of channels c in the convolutional layers of the model, the threshold used to binarize the nonzero pixels of the foreground object, the weight λ of the loss function, the learning rate of Adam, and the number of training epochs. We tuned these hyperparameters to achieve the best Dice similarity coefficient (DSC) (Dice, 1945) between the segmentation result by binarization and the corresponding ground truth of the validation images.

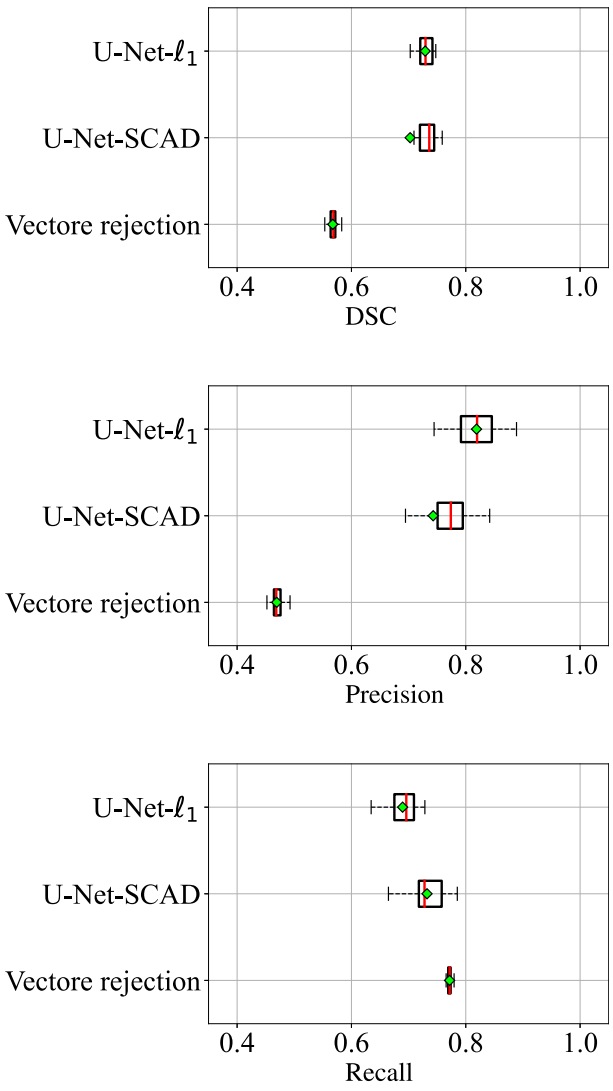


Fig. 2. Box plot of evaluation scores for foreground segmentation results of test images in the "airport" sequence. The red line and green diamond indicate the median and mean scores. The Dice similarity coefficients (DSC) metric represents the harmonic mean of precision and recall, which respectively evaluate over- and under-segmentation. A higher DSC value indicates better segmentation performance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For the RPCA test result, we binarized the vector rejection in Eq. (8) using the matrix of principal components, $\mathbf{L}$, obtained by iterating Eqs. (3), (4) and (5) with $\rho = 1$ on the training images. We determined the best hyperparameters, the binarization threshold and the weight λ in Eq. (2), using Optuna on the validation images in the same manner as with our model hyperparameters. Table 1 presents the hyperparameter values we found.

We performed 20 repetitions of model training and testing, and presented the results of foreground segmentation performance on the test images in Fig. 2. Both U-Net-ℓ_1 and U-Net-SCAD demonstrate similar DSC scores, which are significantly better than those of vector rejection. U-Net-SCAD achieves a balanced precision and recall, resulting in a slight improvement in the detection of missed foreground pixels compared to U-Net-ℓ_1. Vector rejection, on the other hand, shows low precision and is incapable of avoiding numerous false detections from the background.

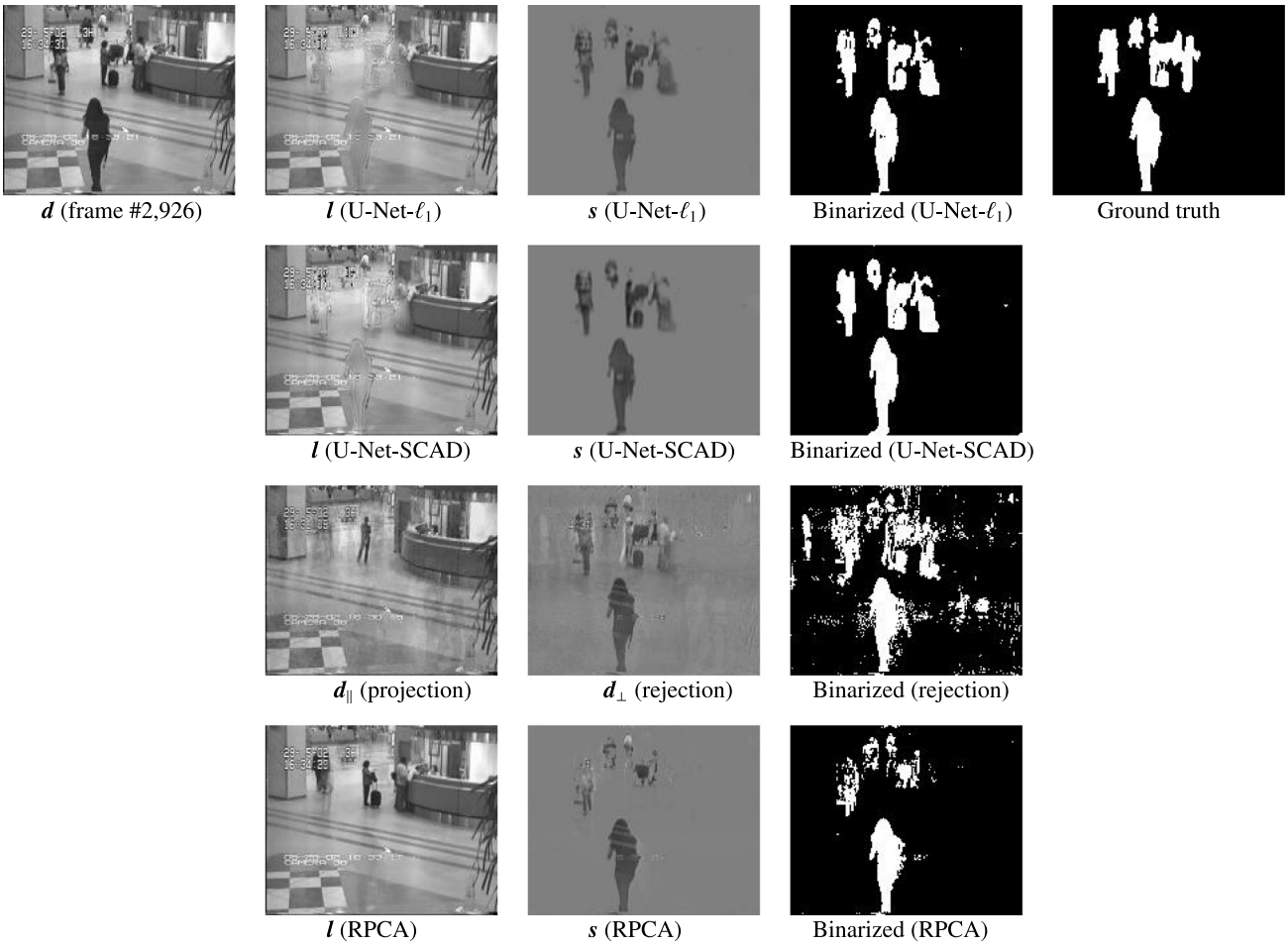


Fig. 3. Example of background subtraction and foreground segmentation of a test image from the "airport" sequence. The top-left panel shows the input image, and the top-right panel displays the ground truth foreground regions (pedestrians). Our models, U-Net-ℓ_1 and U-Net-SCAD, were trained using 50 randomly chosen images from this sequence. We also computed the principal components using the same training set to obtain the vector projection $\boldsymbol{d}_{\|}$ and rejection $\boldsymbol{d}_{\perp}$. Although the test input image (frame #2,926) was not part of the training set, RPCA used 50 successive images including this test frame to produce the results shown in the bottom row.

Fig. 3 shows an example of the background subtraction outcomes for a test image in the sequence. The results demonstrate that our U-Net-ℓ_1 and U-Net-SCAD models identify almost all portions of individuals in the scene as foreground objects represented by $\boldsymbol{s}$, leaving behind only faint silhouettes in the background represented by $\boldsymbol{l} = \boldsymbol{d} - \boldsymbol{s}$. The main difference between the two models lies in the pixel values of $\boldsymbol{s}$. Fig. 4 shows the distributions of the pixel values in the foreground and background regions based on the ground truth. U-Net-ℓ_1 tends to underestimate the absolute values in the foreground regions, while U-Net-SCAD provides high contrast foreground images and therefore the foreground pixels are less likely to be missed.

On the other hand, background subtraction using principal components has several drawbacks in capturing the foreground. The vector projection $\boldsymbol{d}_{\|}$ shows a human silhouette at the upper center of the image, that does not exist in the test image. This person was standing still in many of the training images, and thus was represented by the background's principal components, affecting the vector rejection $\boldsymbol{d}_{\perp} = \boldsymbol{d} - \boldsymbol{d}_{\|}$. Due to the inaccurate subtraction, binary thresholding of the vector rejection $\boldsymbol{d}_{\perp}$ results in unavoidable false detections.

In the bottom row of Fig. 3, we also present the conventional RPCA result obtained by directly applying the ADMM algorithm to a batch of 50 successive frames including this test image. The hyperparameters were adjusted to obtain the best DSC for the test image: The ℓ_1 regularization coefficient λ was 7.7×10^{-3}, and the binarization threshold was 0.08. Despite these adjustments, our models outperforms the RPCA in terms of detecting foreground objects. For instance, the person with a suitcase in the upper center and the person near the pillar in the upper left corner of the test image are represented in the background $\boldsymbol{l}$ by RPCA due to their almost stationary nature, while our models can detect them as foreground $\boldsymbol{s}$. Furthermore, RPCA tends to detect isolated pixels that do not form human shapes. This comparison indicates that our model can learn the apparent shapes of sparse foreground objects without the deed for annotation.

Our model detects pedestrians with similar appearance to those in the training scene. Fig. 5 shows the detection results for the test frame #3,434. Our model could detect almost all pedestrians but the leftmost pedestrian in Fig. 5. This was because the training dataset did not include such a pedestrian with gray appearance. RPCA was also poor at detecting the two leftmost pedestrians because of their small movement. The vector rejection could roughly detect all pedestrians regardless of their appearance, but it suffers from the false positive detection of a spurious human silhouette standing still in background.

An additional benefit of our approach is its computational efficiency. We were able to train the U-Net-ℓ_1 model in just two minutes, using a single NVIDIA Tesla T4 GPU on Google Colaboratory. Moreover, our models have achieved an output rate of around five hundred frames per second, which is about twenty-four times faster than RPCA's $n = 50$ frame output time. As RPCA relies on SVD of $O(mn^2)$ complexity, our U-Net-based models become increasingly advantageous for processing large n frames.

Our approach of learning foreground objects has a clear advantage especially when the background changes are significant. Our U-Net-based models have learned to estimate the foreground, allowing them to generalize well and tolerate realistic changes in camera geometry.

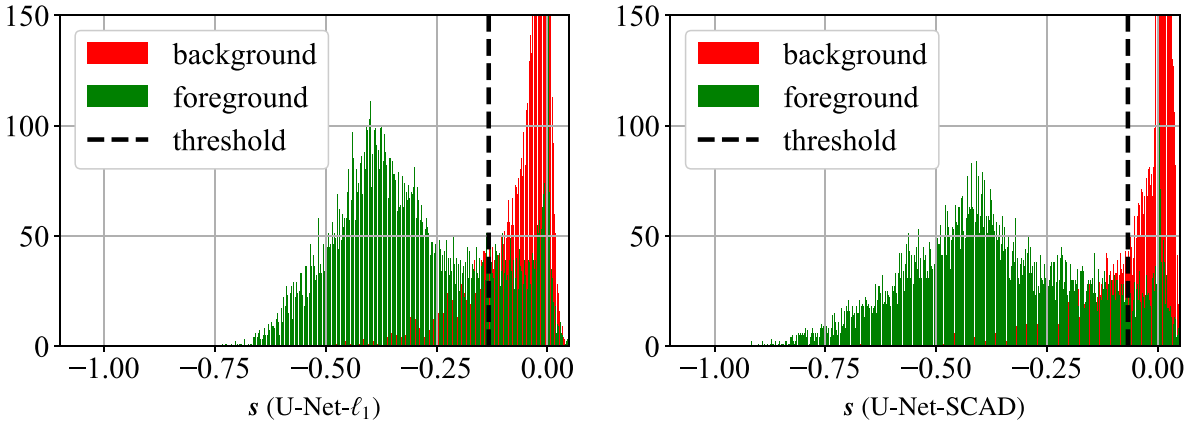


Fig. 4. Distribution of pixel values in the estimated foreground image represented by the model output s for the test input d (frame #2,926). The histograms were made separately for the foreground and background regions based on the ground truth. The threshold for each model output was obtained via hyperparameter tuning, as shown in Table 1.

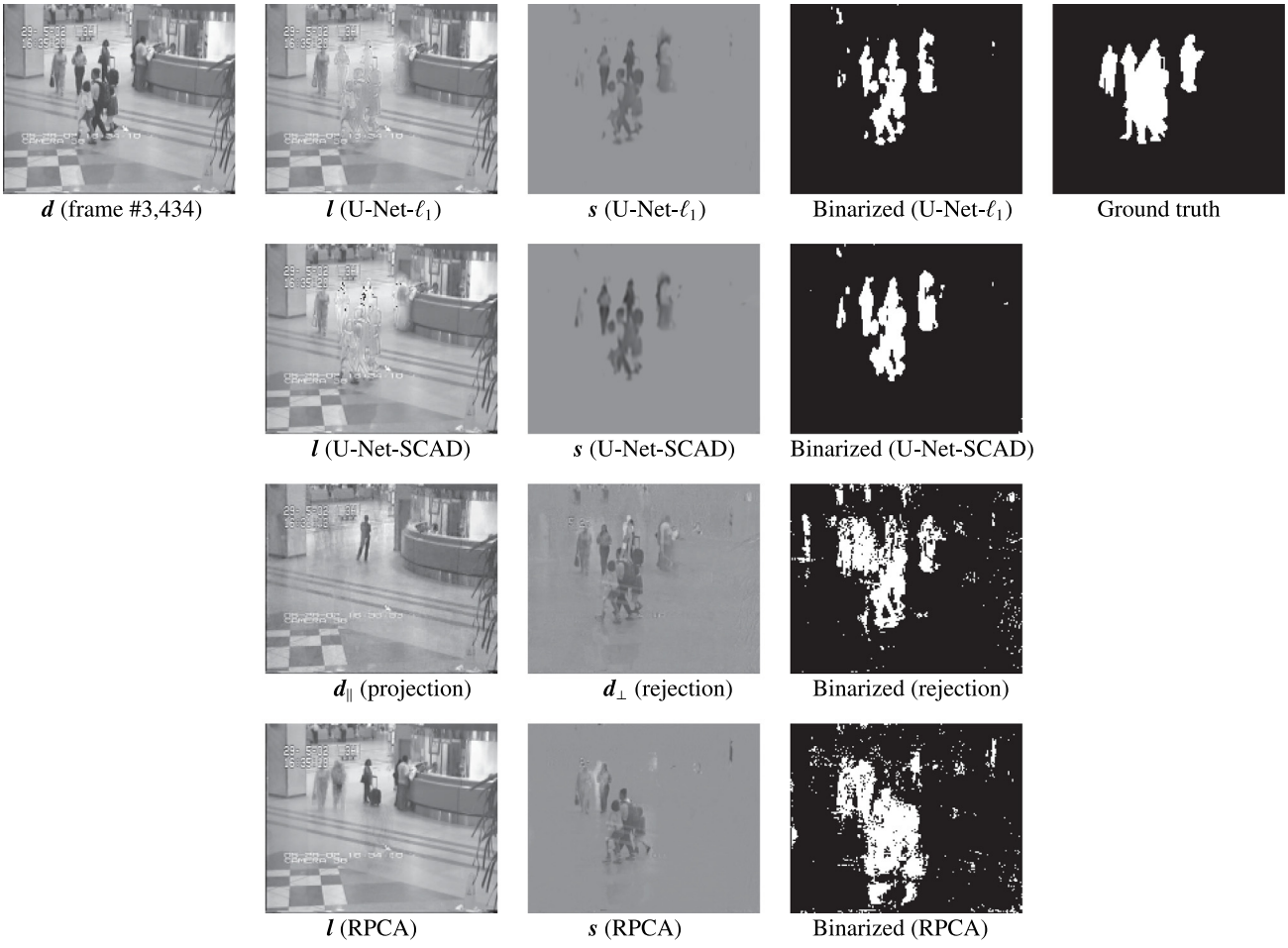


Fig. 5. Same as Fig. 3 for the test input image #3,434.

Table 1
Hyperparameters employed in the background subtraction of "airport" sequence. The number of channels in the convolutional layers are defined by c.

Model	λ	Binarization threshold	c	Learning rate	#epochs
U-Net-ℓ_1 (Eq. (9))	4.3×10^{-3}	0.13	29	4.5×10^{-2}	283
U-Net-SCAD (Eq. (11))	2.3×10^{-2}	0.07	23	2.0×10^{-2}	363
Vector rejection $d_\perp$ (Eq. (8))	1.5×10^{-2}	0.10	–	–	–

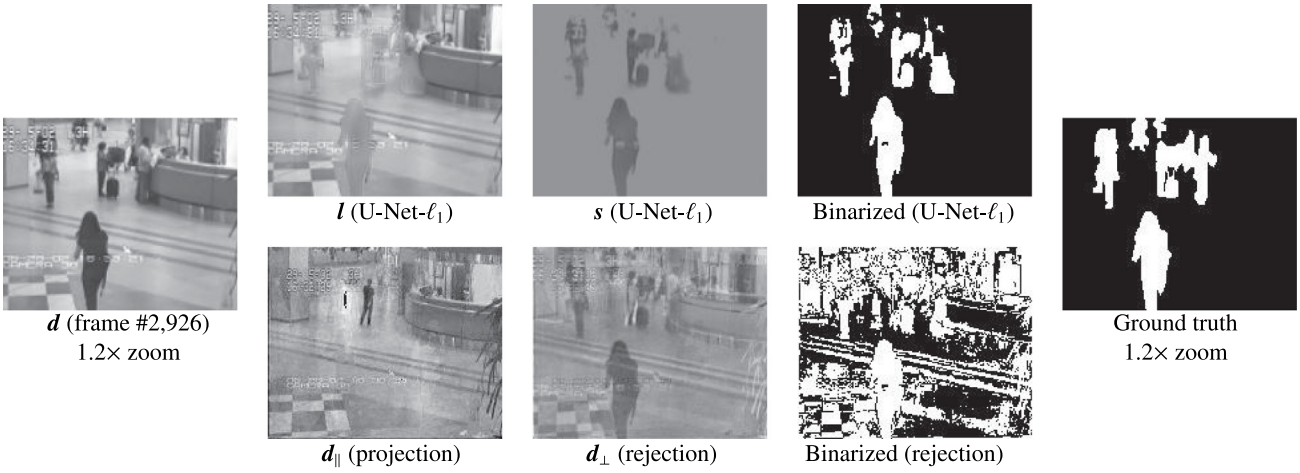


Fig. 6. Background subtraction performed on a test image magnified by a factor of 1.2. The foreground regions were estimated with U-Net-ℓ_1 and vector projection and rejection, following the same techniques as demonstrated in Fig. 3.

Table 2
Hyperparameters used for training two distinct models, namely the unsupervised U-Net-ℓ_1 model and the supervised dual-frame U-Net model, for the purpose of cell segmentation on the ISBI 2014 dataset. The models were trained with different numbers N of training images. The number of channels in the convolutional layers are defined by c.

Model	N	λ	Binarization threshold	c	Learning rate	#epochs
U-Net-ℓ_1	256	1.3×10^{-3}	2.4×10^{-3}	16	5.0×10^{-3}	606
	128	1.4×10^{-3}	2.4×10^{-2}	16	5.0×10^{-3}	690
	64	1.3×10^{-3}	2.4×10^{-2}	16	5.0×10^{-3}	663
	32	1.3×10^{-3}	2.0×10^{-2}	16	5.0×10^{-3}	909
Supervised	256	–	0.5 (fixed)	16	5.0×10^{-3}	744
	128	–	0.5 (fixed)	16	5.0×10^{-3}	744
	64	–	0.5 (fixed)	16	5.0×10^{-3}	643
	32	–	0.5 (fixed)	16	5.0×10^{-3}	452

As demonstrated in the first row of Fig. 6, when presented with a slightly zoomed input image, the U-Net-ℓ_1 model robustly detects the foreground (DSC: 81%, precision: 91%, recall: 74%). In contrast, the principal components of the original image sequence fail to represent the background properly in the zoomed input image, leading to a significant number of false-positive foreground pixels estimated from the vector rejection $\boldsymbol{d}_\perp$ (DSC: 35%, precision: 22%, recall: 81%).

4.2. Learning with small numbers of images

Segmenting cell and nuclei in microscopic images is a fundamental task in biological and biomedical image processing. While deep learning-driven techniques (Ronneberger et al., 2015; Zhou et al., 2018; Jha et al., 2020) have displayed remarkable potential for automatic segmentation, they necessitate substantial amounts of painstakingly labeled training data utilizing pixel-by-pixel annotations by medical experts in supervised learning frameworks.

We present an extensive analysis of our U-Net-ℓ_1 model for unsupervised cell segmentation, investigating the impact of the number of training images on model performance. Our evaluation is conducted on the widely used ISBI 2014 dataset (Lu et al., 2015, 2016), which is a standard benchmark for cell segmentation in cervical cancer cytology. This dataset comprises 945 synthetic grayscale cytology images of size 512×512, each annotated with segmentation masks. For the purpose of this experiment, we modified the original dataset configuration by employing the 45 training images as a validation set for hyperparameter tuning, and the 90 validation images as a test set to assess segmentation quality with the Dice similarity coefficient (DSC). A subset of images from the original 810 test images was used as a training set. For statistical evaluation, we randomly sampled N training images 10 times with N ranging from 32 to 256. We normalized all pixel values to be in a range $[0, 1]$ by simply divided by 255. On the training phase, brightness of each training image were randomly augmented from 60% to 100%.

By leveraging the low-rank and sparse properties of the dataset, we built an unsupervised U-Net-ℓ_1 model by minimizing the loss function as presented in Eq. (9) on the training images. For comparison, we also built a supervised U-Net model with the same architecture as the dual-frame U-Net part of the model in Fig. 1. This supervised model predicts a 512×512 probability map of cell regions from an input image. Instead of nuclear and ℓ_1 norms, the supervised U-Net model was trained with the binary cross-entropy (BCE) loss between the model output and the ground truth of cell segmentation in the training images. We used the Adam optimizer (Kingma and Ba, 2015) with a batch size of 32. The other hyperparameters were tuned by Optuna (Akiba et al., 2019) to maximize the DSC on the validation set. The best values of the hyperparameters for each size N of training set are shown in Table 2.

Fig. 7 shows the test DSC scores of the unsupervised U-Net-ℓ_1 and supervised U-Net models with respect to the number of training images used. Without any supervision, the U-Net-ℓ_1 model achieves the DSC scores close to those of the supervised U-Net model. The low-rank and sparse priors have potential to provide comparable information to annotated large datasets.

We show some examples of cell segmentation test results in Fig. 8. The U-Net-ℓ_1 model trained with a small dataset tends to miss some parts of cell regions with less features, which implies the model distinguishes by learned cell features. The supervised model, even trained with a larger dataset, misidentifies a few faint spots in the background that do not resemble cells. Since it is easy to determine the foreground regions because of not being plain like the background, the supervised model may not have actively learned the foreground cell features even though the DSC scores are high.

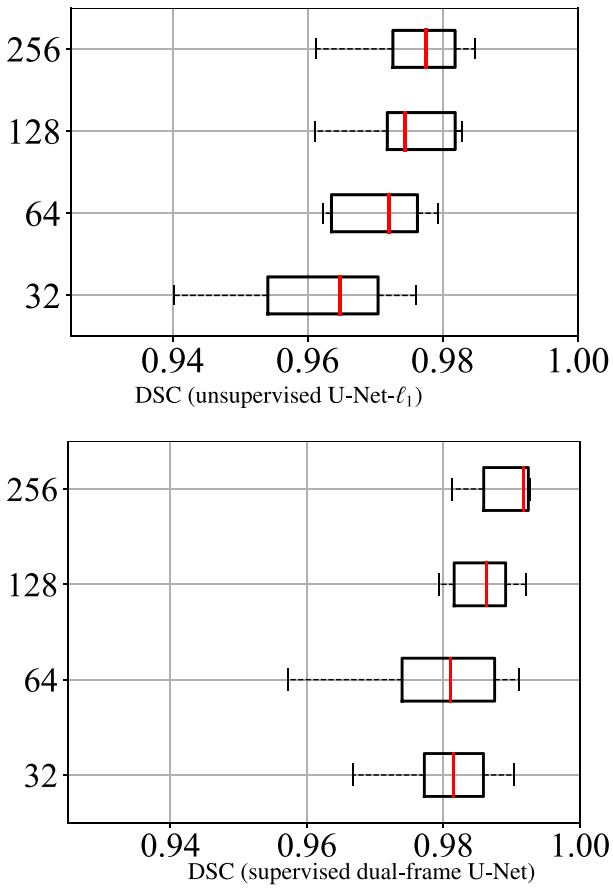


Fig. 7. Box plots of the Dice similarity coefficients (DSCs) for test images with different numbers of training images. The red lines represent the median scores. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We would like to note that the unsupervised model could always provide stable test scores, while the supervised model suffered from severe overfitting without the data augmentation on brightness. The same would obviously apply to any supervised method with larger architecture, making it more hard to avoid overfitting. While the supervised U-Net model was trained to output binary information indicating foreground or background at each pixel, the U-Net-ℓ_1 model is trained as a regression model to encode and decode foreground objects. We conjecture this promotes the U-Net-ℓ_1 model to learn image features more quantitatively from a training set.

4.3. Semantic segmentation of diverse foregrounds

We evaluated the performance of our model in automatically segmenting diverse foreground objects in both grayscale and color images by using the BBBC038v1 image dataset from the Broad Bioimage Benchmark Collection (Caicedo et al., 2019). This image dataset consists of grayscale and color biological images of tens of thousands of nuclei from different organisms treated and imaged under a variety of conditions, such as fluorescent and histologic staining, different magnifications and illuminations.

We instantiated the U-Net-ℓ_1 model shown in Fig. 1 to accept an input image with three color channels. Each input image is preprocessed by resizing to 256 × 256 pixels. All pixel values are normalized by dividing by 255. If bright pixels are dominant in an image, the colors are inverted to have a dark background. All 670 images were randomly split into 512 images for training, 88 images for validation, and 70 images for testing. We set the number of channels $c = 17$ in the convolutional layers and used a ℓ_1 loss weight of $\lambda = 4.1 \times 10^{-3}$. We trained the model for 192 epochs using Adam with a learning of 1.3×10^{-2}. The binarization threshold for foreground segmentation was set to 3.1×10^{-3}. We utilized Optuna (Akiba et al., 2019) to optimize all hyperparameters for maximizing the Dice similarity coefficient (DSC) on the validation set, excluding the batch size of 16 due to GPU memory constraints. It took only about four minutes to train the model on an RTX 3090 GPU. Throughout the optimization process, the training loss and validation loss were observed to decrease almost monotonically until convergence.

On the test images, our unsupervised model achieved the average DSC score of 84% with a standard deviation of 13%. The segmentation results for the test images are illustrated in Fig. 9. As most of the sparse foreground objects in the images are nuclei, it can be concluded that our model has learned their shape and staining features sufficiently from the training images. Our model effectively captures nuclei regions in the test images, including those with less sparsity, multiple contrasts, or low contrast, as depicted in Fig. 9(a), (b), and (c). Overall, the unsupervised learning based only on the low-rank and sparse priors enables us to develop a moderately performing model without any supervision.

The low DSC score for the image in (d) indicates that the model has learned to find foreground cells rather than nuclei inside them. Our model also has another limitation in addressing pathological images such as the one shown in (e) because the foreground nuclei and backgrounds in a set of pathological images do not exhibit L+S structure. For these situations, it is vital to incorporate the ground truth information of cell nuclei. As demonstrated in the latest paper (Tomar et al., 2023), supervised models achieved greater than 90% of F1 score,[1] e.g., 92% by FANet, although they have some drawbacks: expensive annotation cost, intricate model architecture, larger computational demands, and the risk of overfitting.

5. Conclusions

We have demonstrated that incorporating prior knowledge of the linear dependence of training image backgrounds enables a deep neural network to recognize sparse foreground objects. In the proposed framework, a U-Net-based model is trained to encode and decode sparse foreground objects in batches of input images with low-rank backgrounds by minimizing a sparsity-inducing norm together with the nuclear norm.

Our unsupervised deep learning approach offers a more practical and effective alternative to RPCA for capturing image features of interest. Once the U-Net model has learned enough features common to the foreground objects, it can robustly detect them from any single image regardless of the low-rankness and sparseness. The U-Net also enables online object segmentation with much less computational expense than that of RPCA. We expect that our approach will help advance the practicality and utility of the L + S model across various applications.

We have confirmed the advantages of our proposed framework in the context of background subtraction in video surveillance. Additionally, we showed that our method can build a well-generalized cell segmentation model from only a few dozen unannotated training images. As the experimental results indicate, our unsupervised deep learning enable the detection of various foreground objects with different appearances, which opens up new possibilities for providing pre-trained models in the context of transfer learning. Further research should be conducted to explore this potential option.

[1] The F1 score of pixel detection is equivalent to the DSC score in the context of image segmentation.

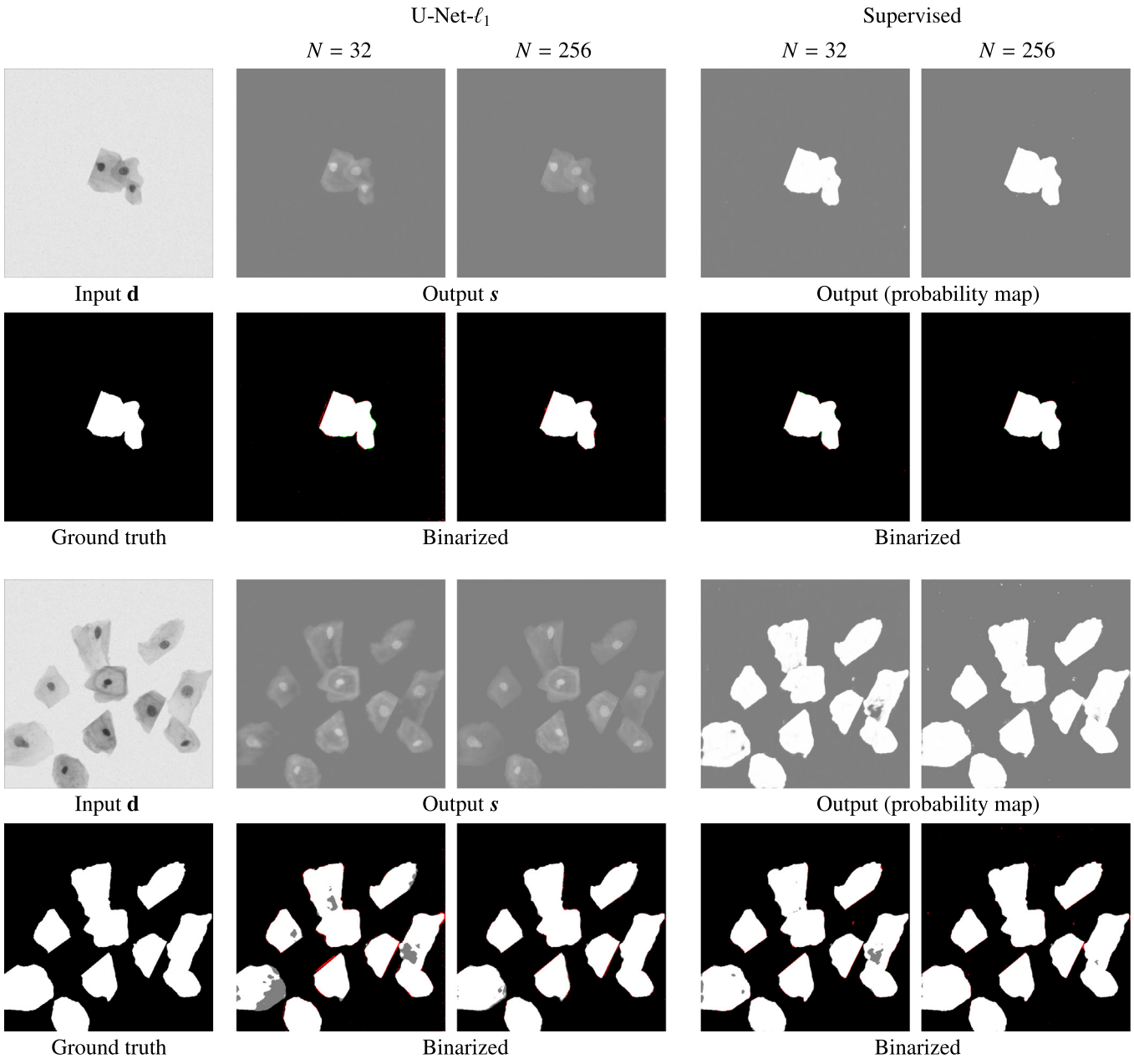


Fig. 8. Examples of segmentation test results on cervical cytology images. The first column displays the input image and ground truth cell regions. The second and third columns show the foreground images predicted by the unsupervised U-Net-ℓ_1 model and their corresponding segmentation results using binarization. Red and gray colors on binarization results indicate false positive and false negative, respectively. The fourth and fifth columns depict the probability maps and their binarization predicted by the supervised U-Net model. The number of images used for training the models is indicated by N. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CRediT authorship contribution statement

Keita Takeda: Formal analysis, Investigation, Methodology, Resources, Software, Validation, Writing – original draft, Writing – review & editing. **Tomoya Sakai:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Tomoya Sakai reports financial support was provided by Japan Society for the Promotion of Science.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP19 H04177 and Nagasaki University Regional Joint Research Support Program 2023.

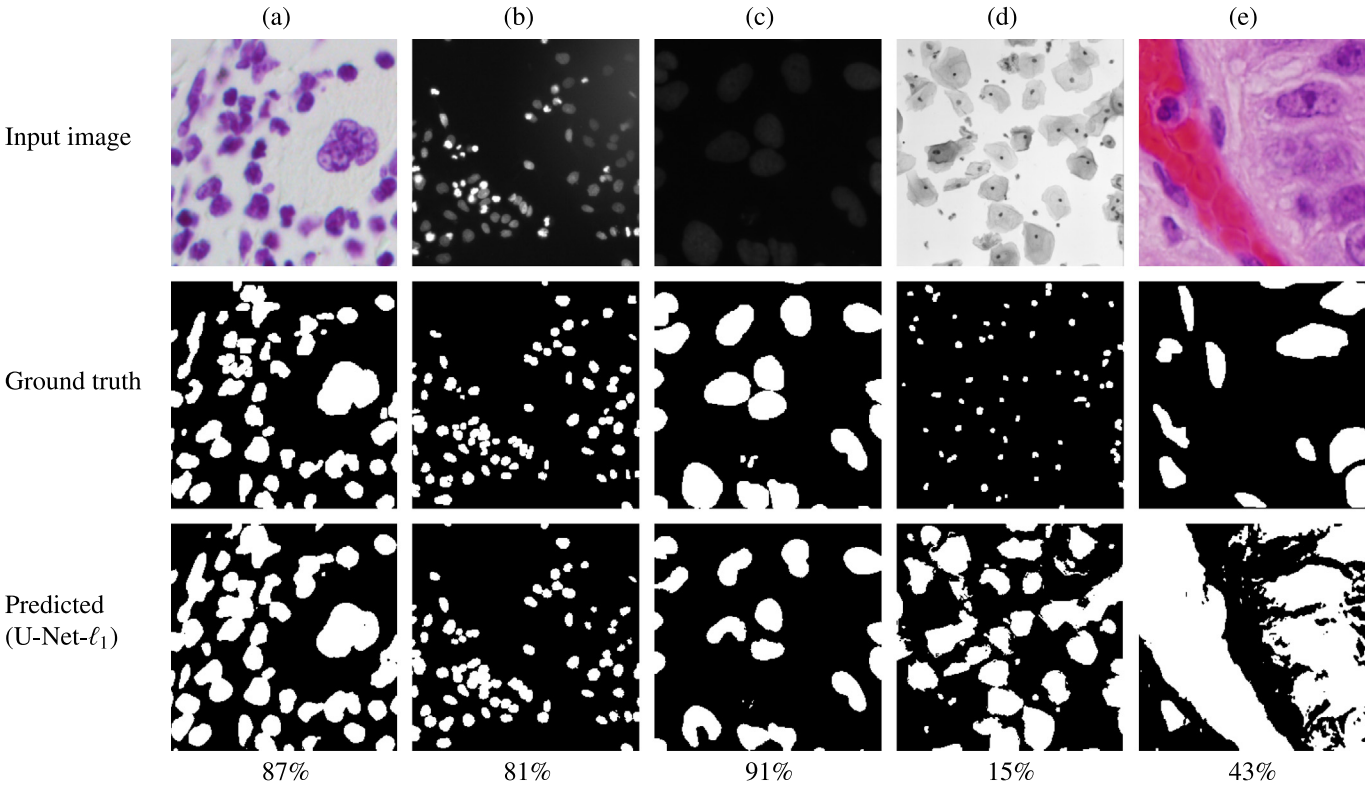


Fig. 9. Evaluation of cell nuclei segmentation performance using the U-Net-ℓ_1 model on the BBBC038v1 image dataset (Caicedo et al., 2019). Segmentation results for color (a), grayscale (b), and low-contrast (c) images are shown, along with images featuring a small nucleus relative to the cytoplasm (d) and a pathology image (e). The Dice similarity coefficient, representing the agreement between predicted and ground truth foregrounds, is displayed as a percentage at the bottom of each column.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2623–2631.

Bouwmans, T., Zahzah, E.H., 2014. Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance. Comput. Vis. Image Underst. 122, 22–34.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. 3 (1), 1–122.

Cai, J.-F., Candès, E.J., Shen, Z., 2010. A singular value thresholding algorithm for matrix completion. SIAM J. Optim. 20 (4), 1956–1982.

Cai, H., Liu, J., Yin, W., 2021. Learned robust PCA: A scalable deep unfolding approach for high-dimensional outlier detection. Adv. Neural Inf. Process. Syst. 34, 16977–16989.

Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., et al., 2019. Nucleus segmentation across imaging experiments: The 2018 Data Science Bowl. Nature Methods 16 (12), 1247–1253.

Candès, E.J., Li, X., Ma, Y., Wright, J., 2011. Robust principal component analysis? J. ACM 58 (3), 11:1–11:37.

Daubechies, I., Defrise, M., De Mol, C., 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm. Pure Appl. Math. 57 (11), 1413–1457.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Donoho, D.L., 1995. De-noising by soft-thresholding. Inf. Theory IEEE Trans. on 41 (3), 613–627.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96 (456), 1348–1360.

Gabay, D., Mercier, B., 1976. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. 2 (1), 17–40.

Guyon, C., Bouwmans, T., Zahzah, E.-h., et al., 2012. Robust principal component analysis for background subtraction: Systematic evaluation and comparative analysis. Principal Component Anal. 10, 223–238.

Han, Y., Ye, J.C., 2018. Framing U-Net via deep convolutional framelets: Application to sparse-view CT. IEEE Trans. Med. Imaging 37 (6), 1418–1429.

Huang, P.-S., Chen, S.D., Smaragdis, P., Hasegawa-Johnson, M., 2012. Singing-voice separation from monaural recordings using robust principal component analysis. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 57–60.

Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P., Johansen, H.D., 2020. DoubleU-Net: A deep convolutional neural network for medical image segmentation. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems. IEEE, pp. 558–564.

Kalsotra, R., Arora, S., 2022. Background subtraction for moving object detection: explorations of recent developments and challenges. Vis. Comput. 38 (12), 4151–4178.

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

Li, L., Huang, W., Gu, I.Y.-H., Tian, Q., 2004. Statistical modeling of complex backgrounds for foreground object detection. IEEE Trans. Image Process. 13 (11), 1459–1472.

Lu, Z., Carneiro, G., Bradley, A.P., 2015. An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells. IEEE Trans. Image Process. 24 (4), 1261–1272.

Lu, Z., Carneiro, G., Bradley, A.P., Ushizima, D., Nosrati, M.S., Bianchi, A.G., Carneiro, C.M., Hamarneh, G., 2016. Evaluation of three algorithms for the segmentation of overlapping cervical cells. IEEE J. Biomed. Health Inf. 21 (2), 441–450.

Ma, S., Goldfarb, D., Chen, L., 2011. Fixed point and Bregman iterative methods for matrix rank minimization. Math. Program. 128 (1–2), 321–353.

Paszke, A., et al., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), In: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., pp. 8026–8037.

Rezaei, B., Farnoosh, A., Ostadabbas, S., 2020. G-LBM: Generative low-dimensional background model estimation from video sequences. In: European Conference on Computer Vision. Springer, pp. 293–310.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Sakai, T., Kuhara, H., 2015. Separating background and foreground optical flow fields by low-rank and sparse regularization. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, pp. 1523–1527.

Sakai, T., Miyahara, S., Kiyasu, S., 2016. Unmixing three types of lung sounds by convex optimization. In: 2016 23rd International Conference on Pattern Recognition. ICPR, IEEE, pp. 2884–2888.

Siddique, N., Paheding, S., Elkin, C.P., Devabhaktuni, V., 2021. U-net and its variants for medical image segmentation: a review of theory and applications. IEEE Access 9, 82031–82057. http://dx.doi.org/10.1109/ACCESS.2021.3086020.

Skočaj, D., Leonardis, A., Bischof, H., 2007. Weighted and robust learning of subspace representations. Pattern Recognit. 40 (5), 1556–1569. http://dx.doi.org/10.1016/j.patcog.2006.09.019, URL: https://www.sciencedirect.com/science/article/pii/S0031320306003876.

Solomon, O., Cohen, R., Zhang, Y., Yang, Y., He, Q., Luo, J., van Sloun, R.J.G., Eldar, Y.C., 2020. Deep unfolded robust PCA with application to clutter suppression in ultrasound. IEEE Trans. Med. Imaging 39, 1051–1067.

Takeda, K., Fujiwara, K., Sakai, T., 2022. Unsupervised deep learning for online foreground segmentation exploiting low-rank and sparse priors. In: 2022 International Conference on Digital Image Computing: Techniques and Applications. DICTA, pp. 1–7. http://dx.doi.org/10.1109/DICTA56598.2022.10034581.

Tomar, N.K., Jha, D., Riegler, M.A., Johansen, H.D., Johansen, D., Rittscher, J., Halvorsen, P., Ali, S., 2023. Fanet: a feedback attention network for improved biomedical image segmentation. IEEE Trans. Neural Netw. Learn. Syst. 34 (11), 9375–9388. http://dx.doi.org/10.1109/TNNLS.2022.3159394.

Zhang, N., Ashikuzzaman, M., Rivaz, H., 2020. Clutter suppression in ultrasound: Performance evaluation and review of low-rank and sparse matrix decomposition methods. BioMed. Eng. OnLine 19 (1).

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. UNet++: A nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 3–11.