# U-Net: Convolutional Networks for Biomedical Image Segmentation

Olaf Ronneberger, Philipp Fischer, and Thomas Brox

Computer Science Department and BIOSS Centre for Biological Signalling Studies,
University of Freiburg, Germany
ronneber@informatik.uni-freiburg.de,
WWW home page: http://lmb.informatik.uni-freiburg.de/

**Abstract.** There is large consent that successful training of deep networks requires many thousand annotated training samples. In this paper, we present a network and training strategy that relies on the strong use of data augmentation to use the available annotated samples more efficiently. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. We show that such a network can be trained end-to-end from very few images and outperforms the prior best method (a sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. Using the same network trained on transmitted light microscopy images (phase contrast and DIC) we won the ISBI cell tracking challenge 2015 in these categories by a large margin. Moreover, the network is fast. Segmentation of a 512x512 image takes less than a second on a recent GPU. The full implementation (based on Caffe) and the trained networks are available at http://lmb.informatik.uni-freiburg.de/people/ronneber/u-net.

## 1 Introduction

In the last two years, deep convolutional networks have outperformed the state of the art in many visual recognition tasks, e.g. [7,3]. While convolutional networks have already existed for a long time [8], their success was limited due to the size of the available training sets and the size of the considered networks. The breakthrough by Krizhevsky et al. [7] was due to supervised training of a large network with 8 layers and millions of parameters on the ImageNet dataset with 1 million training images. Since then, even larger and deeper networks have been trained [12].

The typical use of convolutional networks is on classification tasks, where the output to an image is a single class label. However, in many visual tasks, especially in biomedical image processing, the desired output should include localization, i.e., a class label is supposed to be assigned to each pixel. Moreover, thousands of training images are usually beyond reach in biomedical tasks. Hence, Ciresan et al. [1] trained a network in a sliding-window setup to predict the class label of each pixel by providing a local region (patch) around that pixel

# U-Net: Convolutional Networks for Biomedical Image Segmentation

Olaf Ronneberger, Philipp Fischer và Thomas Brox

Khoa Khoa học Máy tính và Trung tâm nghiên cứu tín hiệu sinh học, Đại học Freiburg, Đức ronneber@informatik.uni-freiburg.de, Trang chủ www: http://lmb.informatik.uni-freiburg.de/

**Abstract.** Có sự đồng ý lớn rằng đào tạo thành công các công trình net sâu yêu cầu nhiều ngàn mẫu đào tạo chú thích. Trong phần này, chúng tôi trình bày một mạng lưới và chiến lược đào tạo phụ thuộc vào việc sử dụng mạnh mẽ việc tăng dữ liệu để sử dụng các mẫu được chú thích có sẵn một cách hiệu quả hơn. Kiến trúc bao gồm một con đường hợp đồng để nắm bắt bối cảnh và một con đường mở rộng đối xứng cho phép địa phương chính xác. Chúng tôi cho thấy rằng một mạng như vậy có thể được đào tạo từ đầu đến cuối từ rất ít hình ảnh và vượt trội so với phương pháp tốt nhất trước đó (mạng chập cửa sổ trượt) trên Thách thức ISBI để phân đoạn các cấu trúc neu trong các ngăn xếp kính hiển vi điện tử. Sử dụng cùng một công việc mạng được đào tạo trên các hình ảnh kính hiển vi ánh sáng truyền qua (độ tương phản pha và DIC), chúng tôi đã giành được Thử thách theo dõi tế bào ISBI 2015 trong các cấu trúc này bằng một biên độ lớn. Hơn nữa, mạng là nhanh. Phân đoạn của hình ảnh 512x512 mất ít hơn một giây trên GPU gần đây. Việc triển khai đầy đủ (dựa trên CA ff e) và các mạng được đào tạo có sẵn tại http://lmb.informatik.uni-freiburg.de/people/ronneber/u-net.

## 1 Introduction

Trong hai năm qua, các mạng tích chập sâu đã vượt trội so với trạng thái của nghệ thuật trong nhiều nhiệm vụ nhận dạng trực quan, ví dụ: [7,3]. Mặc dù các mạng tích chập đã tồn tại trong một thời gian dài [8], thành công của chúng bị hạn chế do quy mô của các bộ đào tạo có sẵn và quy mô của các mạng được xem xét. Bước đột phá của Krizhevsky et al. [7] là do đào tạo có giám sát một mạng lưới lớn với 8 lớp và hàng triệu thông số trên bộ dữ liệu ImageNet với 1 triệu hình ảnh đào tạo. Kể từ đó, thậm chí các mạng lớn hơn và sâu hơn đã được đào tạo [12].

Việc sử dụng điển hình của các mạng tích chập nằm trong các tác vụ phân loại, trong đó đầu ra cho hình ảnh là một nhãn lớp đơn. Tuy nhiên, trong nhiều tác vụ trực quan, đặc biệt là trong xử lý hình ảnh y sinh, đầu ra mong muốn nên bao gồm nội địa hóa, tức là, một nhãn lớp được cho là được gán cho mỗi pixel. Hơn nữa, hàng ngàn hình ảnh đào tạo thường nằm ngoài tầm với trong các nhiệm vụ y sinh. Do đó, Ciresan et al. [1] đã đào tạo một mạng trong thiết lập cửa sổ trượt để dự đoán nhãn lớp của mỗi pixel bằng cách cung cấp một vùng cục bộ (bản vá)

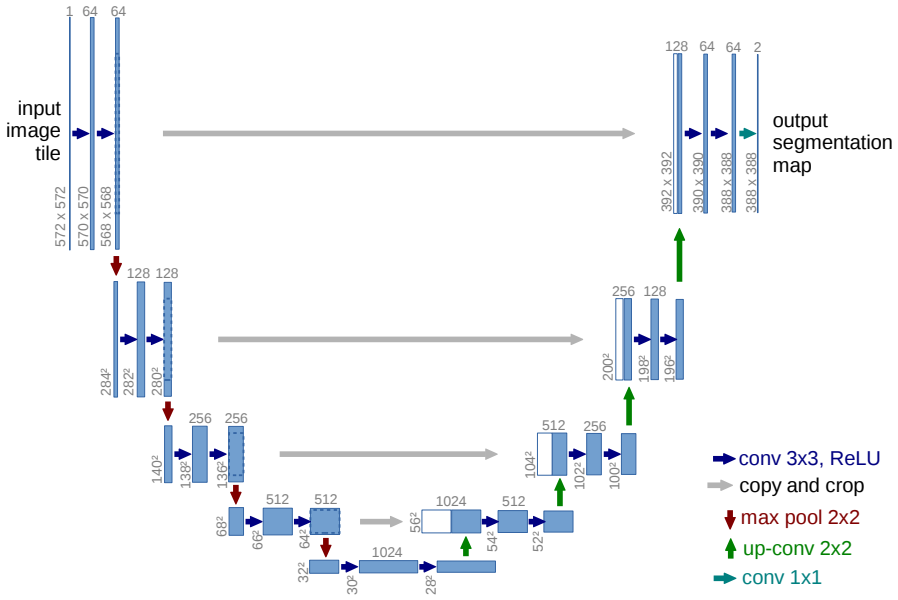


**Fig. 1.** U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

as input. First, this network can localize. Secondly, the training data in terms of patches is much larger than the number of training images. The resulting network won the EM segmentation challenge at ISBI 2012 by a large margin.

Obviously, the strategy in Ciresan et al. [1] has two drawbacks. First, it is quite slow because the network must be run separately for each patch, and there is a lot of redundancy due to overlapping patches. Secondly, there is a trade-off between localization accuracy and the use of context. Larger patches require more max-pooling layers that reduce the localization accuracy, while small patches allow the network to see only little context. More recent approaches [11,4] proposed a classifier output that takes into account the features from multiple layers. Good localization and the use of context are possible at the same time.

In this paper, we build upon a more elegant architecture, the so-called "fully convolutional network" [9]. We modify and extend this architecture such that it works with very few training images and yields more precise segmentations; see Figure 1. The main idea in [9] is to supplement a usual contracting network by successive layers, where pooling operators are replaced by upsampling operators. Hence, these layers increase the resolution of the output. In order to localize, high resolution features from the contracting path are combined with the upsampled

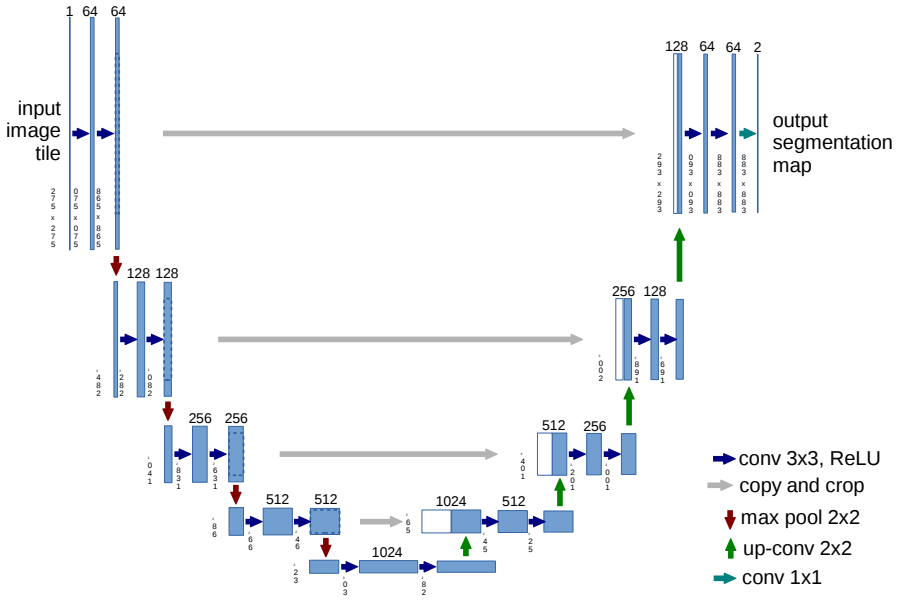


**Fig. 1.** Kiến trúc U-NET (ví dụ cho 32x32 pixel ở độ phân giải thấp nhất). Mỗi hộp màu xan h tương ứng với bản đồ tính năng đa kênh. Số lượng kênh được biểu thị trên đầu hộp. Kích thước X-Y được cung cấp ở cạnh dưới bên trái của hộp. Hộp trắng đại diện cho bản đồ tín h năng sao chép. Các mũi tên biểu thị các hoạt động khác nhau.

như đầu vào. Đầu tiên, mạng này có thể bản địa hóa. Thứ hai, dữ liệu đào tạo về c ác bản vá lớn hơn nhiều so với số lượng hình ảnh đào tạo. Mạng kết quả đã giành được Thử thách phân khúc EM tại ISBI 2012 bằng một biên độ lớn.

Rõ ràng, chiến lược trong Ciresan et al. [1] có hai nhược điểm. Đầu tiên, nó khá chậm vì mạng phải được chạy riêng cho mỗi bản vá và có rất nhiều dự phòng do cá c bản vá chồng chéo. Thứ hai, có một sự thương mại giữa độ chính xác nội địa hóa và việc sử dụng bối cảnh. Các bản vá lớn hơn yêu cầu các lớp nhóm tối đa hơn làm giảm độ chính xác nội địa hóa, trong khi các bản vá nhỏ cho phép mạng chỉ thấy ít b ối cảnh. Các cách tiếp cận gần đây hơn [11,4] đã đề xuất một đầu ra phân loại có tí nh đến các tính năng từ nhiều lớp. Nội địa hóa tốt và việc sử dụng bối cảnh là có t hể cùng một lúc.

Trong bài báo này, chúng tôi xây dựng dựa trên một kiến trúc thanh lịch hơn, cái gọi là mạng tích chập hoàn toàn của Google [9]. Chúng tôi sửa đổi và mở rộng kiến trúc này sao cho nó hoạt động với rất ít hình ảnh đào tạo và mang lại các phân đoạn chính xác hơn; Xem Hình 1. Ý tưởng chính trong [9] là bổ sung một mạng hợp đồng thông thường bằng các lớp liên tiếp, trong đó các toán tử gộp được thay thế bằng c ác toán tử lấy mẫu. Do đó, các lớp này làm tăng độ phân giải của đầu ra. Để bản đị a hóa, các tính năng có độ phân giải cao từ đường dẫn hợp đồng được kết hợp với UPSampled

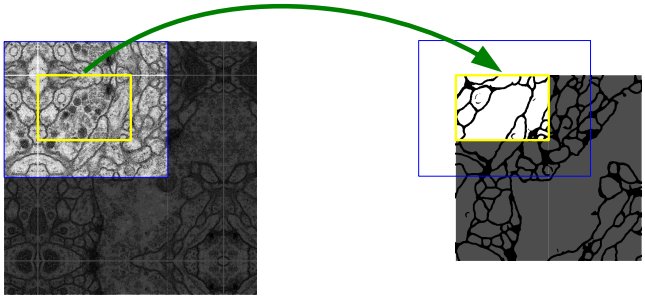

**Fig. 2.** Overlap-tile strategy for seamless segmentation of arbitrary large images (here segmentation of neuronal structures in EM stacks). Prediction of the segmentation in the yellow area, requires image data within the blue area as input. Missing input data is extrapolated by mirroring

output. A successive convolution layer can then learn to assemble a more precise output based on this information.

One important modification in our architecture is that in the upsampling part we have also a large number of feature channels, which allow the network to propagate context information to higher resolution layers. As a consequence, the expansive path is more or less symmetric to the contracting path, and yields a u-shaped architecture. The network does not have any fully connected layers and only uses the valid part of each convolution, i.e., the segmentation map only contains the pixels, for which the full context is available in the input image. This strategy allows the seamless segmentation of arbitrarily large images by an overlap-tile strategy (see Figure 2). To predict the pixels in the border region of the image, the missing context is extrapolated by mirroring the input image. This tiling strategy is important to apply the network to large images, since otherwise the resolution would be limited by the GPU memory.

As for our tasks there is very little training data available, we use excessive data augmentation by applying elastic deformations to the available training images. This allows the network to learn invariance to such deformations, without the need to see these transformations in the annotated image corpus. This is particularly important in biomedical segmentation, since deformation used to be the most common variation in tissue and realistic deformations can be simulated efficiently. The value of data augmentation for learning invariance has been shown in Dosovitskiy et al. [2] in the scope of unsupervised feature learning.

Another challenge in many cell segmentation tasks is the separation of touching objects of the same class; see Figure 3. To this end, we propose the use of a weighted loss, where the separating background labels between touching cells obtain a large weight in the loss function.

The resulting network is applicable to various biomedical segmentation problems. In this paper, we show results on the segmentation of neuronal structures in EM stacks (an ongoing competition started at ISBI 2012), where we out-

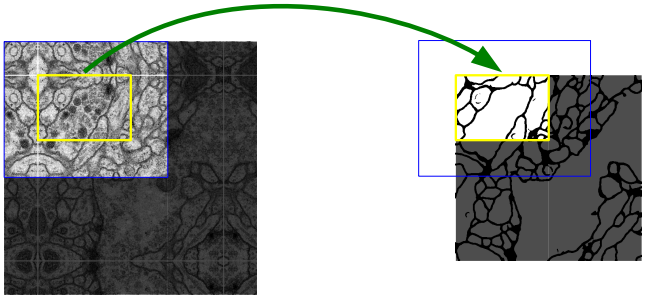

**Fig. 2.** Chiến lược gạch chồng chéo để phân đoạn liền mạch các hình ảnh lớn tùy ý (ở đây phân đoạn cấu trúc tế bào thần kinh trong các ngăn xếp EM). Dự đoán phân đoạn trong khu vực màu vàng, yêu cầu dữ liệu hình ảnh trong khu vực màu xanh làm đầu vào. Thiếu dữ liệu đầu vào được ngoại suy bằng cách phản chiếu

đầu ra. Một lớp tích chập liên tiếp sau đó có thể học cách lắp ráp một đầu ra chính xác hơn dựa trên thông tin này.

Một điều chỉnh quan trọng trong kiến trúc của chúng tôi là trong phần UPSampli ng, chúng tôi cũng có một số lượng lớn các kênh tính năng, cho phép mạng truyền b á thông tin ngữ cảnh đến các lớp độ phân giải cao hơn. Kết quả là, con đường mở rộng ít nhiều đối xứng với đường dẫn hợp đồng và mang lại một kiến trúc hình ch ữ U. Mạng không có bất kỳ lớp kết nối đầy đủ nào và chỉ sử dụng phần hợp lệ của mỗi lần chập, tức là bản đồ phân đoạn chỉ chứa các pixel, trong đó bối cảnh đầy đ ủ có sẵn trong hình ảnh đầu vào. Chiến lược này cho phép phân đoạn liền mạch củ a các hình ảnh lớn tùy ý bằng một chiến lược gạch chồng chéo (xem Hình 2). Để d ự đoán các pixel trong vùng viền của hình ảnh, bối cảnh bị thiếu được ngoại suy bằ ng cách phản chiếu hình ảnh đầu vào. Chiến lược ốp lát này rất quan trọng để áp d ụng mạng vào các hình ảnh lớn, vì nếu không độ phân giải sẽ bị giới hạn bởi bộ nh ớ GPU.

Đối với các nhiệm vụ của chúng tôi, có rất ít dữ liệu đào tạo có sẵn, chúng tôi s ử dụng tăng dữ liệu quá mức bằng cách áp dụng các biến dạng đàn hồi cho các thờ i gian đào tạo có sẵn. Điều này cho phép mạng học được bất biến đối với các biến dạng như vậy, mà không cần phải xem các biến đổi này trong kho chứa hình ảnh đ ược chú thích. Điều này đặc biệt quan trọng trong phân đoạn y sinh, vì biến dạng đ ược sử dụng là biến thể phổ biến nhất trong mô và biến dạng thực tế có thể được mô phỏng một cách hiệu quả. Giá trị của việc tăng dữ liệu cho sự bất biến học tập đã được thể hiện trong Dosovitskiy et al. [2] trong phạm vi học tập tính năng không giám sát.

Một thách thức khác trong nhiều nhiệm vụ phân đoạn tế bào là tách các đối tượ ng cảm ứng của cùng một lớp; Xem Hình 3. Đến đầu này, chúng tôi đề xuất việc s ử dụng một sự mất trọng có trọng số, trong đó các nhãn nền tách biệt giữa các ô ch ạm có trọng lượng lớn trong hàm mất.

Mạng kết quả được áp dụng cho các phân đoạn y sinh khác nhau. Trong bài báo này, chúng tôi cho thấy kết quả về phân đoạn cấu trúc tế bào thần kinh trong các ng ăn xếp EM (một cuộc thi đang diễn ra bắt đầu tại ISBI 2012), nơi chúng tôi ra ngoà i-

performed the network of Ciresan et al. [1]. Furthermore, we show results for cell segmentation in light microscopy images from the ISBI cell tracking challenge 2015. Here we won with a large margin on the two most challenging 2D transmitted light datasets.

## 2 Network Architecture

The network architecture is illustrated in Figure 1. It consists of a contracting path (left side) and an expansive path (right side). The contracting path follows the typical architecture of a convolutional network. It consists of the repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step we double the number of feature channels. Every step in the expansive path consists of an upsampling of the feature map followed by a 2x2 convolution ("up-convolution") that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions, each followed by a ReLU. The cropping is necessary due to the loss of border pixels in every convolution. At the final layer a 1x1 convolution is used to map each 64-component feature vector to the desired number of classes. In total the network has 23 convolutional layers.

To allow a seamless tiling of the output segmentation map (see Figure 2), it is important to select the input tile size such that all 2x2 max-pooling operations are applied to a layer with an even x- and y-size.

## 3 Training

The input images and their corresponding segmentation maps are used to train the network with the stochastic gradient descent implementation of Caffe [6]. Due to the unpadded convolutions, the output image is smaller than the input by a constant border width. To minimize the overhead and make maximum use of the GPU memory, we favor large input tiles over a large batch size and hence reduce the batch to a single image. Accordingly we use a high momentum (0.99) such that a large number of the previously seen training samples determine the update in the current optimization step.

The energy function is computed by a pixel-wise soft-max over the final feature map combined with the cross entropy loss function. The soft-max is defined as $p_k(\mathbf{x}) = \exp(a_k(\mathbf{x}))/\left(\sum_{k'=1}^{K}\exp(a_{k'}(\mathbf{x}))\right)$ where $a_k(\mathbf{x})$ denotes the activation in feature channel $k$ at the pixel position $\mathbf{x} \in \Omega$ with $\Omega \subset \mathbb{Z}^2$. $K$ is the number of classes and $p_k(\mathbf{x})$ is the approximated maximum-function. I.e. $p_k(\mathbf{x}) \approx 1$ for the $k$ that has the maximum activation $a_k(\mathbf{x})$ and $p_k(\mathbf{x}) \approx 0$ for all other $k$. The cross entropy then penalizes at each position the deviation of $p_{\ell(\mathbf{x})}(\mathbf{x})$ from 1 using

$$E = \sum_{\mathbf{x}\in\Omega} w(\mathbf{x})\log(p_{\ell(\mathbf{x})}(\mathbf{x})) \tag{1}$$

Thực hiện mạng lưới của Ciresan et al. [1]. Hơn nữa, chúng tôi hiển thị kết quả phân đoạn tế bào trong hình ảnh kính hiển vi ánh sáng từ tế bào ISBI theo dõi Chal-lenge 2015. Ở đây chúng tôi đã giành chiến thắng với một biên độ lớn trên hai bộ dữ liệu ánh sáng 2D thử thách nhất.

## 2 Network Architecture

Kiến trúc mạng được minh họa trong Hình 1. Nó bao gồm một đường dẫn hợp đồng (bên trái) và một đường dẫn mở rộng (bên phải). Đường dẫn hợp đồng tuân theo kiến trúc điển hình của một mạng lưới tích chập. Nó bao gồm ứng dụng lặp đi lặp lại của hai lần chập 3x3 (tích chập không được sử dụng), mỗi lần theo sau là một đơn vị tuyến tính trực tràng (RELU) và hoạt động gộp tối đa 2x2 với Stride 2 để lấy mẫu xuống. Ở mỗi bước Downsampling, chúng tôi tăng gấp đôi số lượng kênh tính năng. Mỗi bước trong đường dẫn mở rộng bao gồm việc lấy mẫu của bản đồ tính năng theo sau là tích chập 2x2 (cách mạng UP), giảm một nửa số lượng kênh tính năng, kết nối với bản đồ tính năng được cắt tương ứng từ đường dẫn hợp đồng và hai lần chập số 3x3, mỗi lần giảm dần. Việc cắt xén là cần thiết do mất các pixel biên giới trong mỗi lần kết hợp. Ở lớp cuối cùng, tích chập 1x1 được sử dụng để ánh xạ từng vectơ tính năng thành phần 64 cho số lượng lớp mong muốn. Tổng cộng mạng có 23 lớp chập.

Để cho phép một ốp lát liền mạch của bản đồ phân đoạn đầu ra (xem Hình 2), điều quan trọng là chọn kích thước gạch đầu vào sao cho tất cả các hoạt động của nhóm tối đa 2x2 được áp dụng cho một lớp có kích thước X và Y chẵn.

## 3 Training

Các hình ảnh đầu vào và các bản đồ phân đoạn tương ứng của chúng được sử dụng để đào tạo mạng với việc thực hiện giảm độ dốc ngẫu nhiên của Caffe [6]. Do các chập chập không được bảo vệ, hình ảnh đầu ra nhỏ hơn đầu vào của chiều rộng đường viền không đổi. Để giảm thiểu chi phí và sử dụng tối đa bộ nhớ GPU, chúng tôi ủng hộ các ô đầu vào lớn trên một kích thước lô lớn và do đó giảm lô xuống một hình ảnh duy nhất. Theo đó, chúng tôi sử dụng động lượng cao (0,99) sao cho một số lượng lớn các mẫu đào tạo được thấy trước đây xác định bản cập nhật trong bước tối ưu hóa hiện tại.

Hàm năng lượng được tính toán bởi một max mềm thông minh trên bản đồ tính năng cuối cùng kết hợp với chức năng mất entropy chéo. Phần mềm mềm được định nghĩa là $p_k(\mathbf{x}) = \exp\left(a_k(\mathbf{x})\right) / \left(\sum_{k'=1}^{K} \exp\left(a_{k'}(\mathbf{x})\right)\right)$ trong đó $a_k(\mathbf{x})$ biểu thị kích hoạt trong kênh tính năng $k$ ở vị trí pixel $\mathbf{x} \in \Omega$ với $\Omega \subset \mathbb{Z}^2$. $K$ là số lớp và $p_k(\mathbf{x})$ là chức năng tối đa gần đúng. Tức là $p_k(\mathbf{x}) \approx 1$ cho $k$ có kích hoạt tối đa $a_k(\mathbf{x})$ và $p_k(\mathbf{x}) \approx 0$ cho tất cả các $k$ khác. Entropy chéo sau đó xử phạt ở mỗi vị trí độ lệch của $p_{\ell(\mathbf{x})}(\mathbf{x})$ từ 1 bằng cách sử dụng

$$E = \sum_{\mathbf{x} \in \Omega} w(\mathbf{x}) \log(p_{\ell(\mathbf{x})}(\mathbf{x})) \tag{1}$$

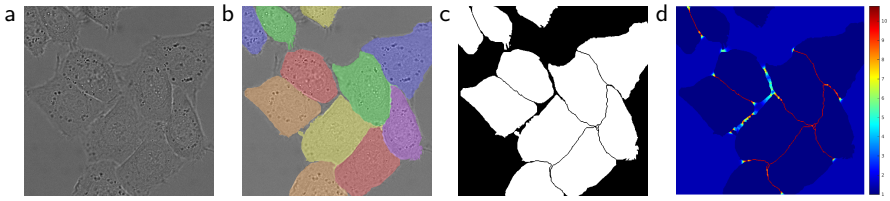


**Fig. 3.** HeLa cells on glass recorded with DIC (differential interference contrast) microscopy. (**a**) raw image. (**b**) overlay with ground truth segmentation. Different colors indicate different instances of the HeLa cells. (**c**) generated segmentation mask (white: foreground, black: background). (**d**) map with a pixel-wise loss weight to force the network to learn the border pixels.

where $\ell : \Omega \to \{1, \dots, K\}$ is the true label of each pixel and $w : \Omega \to \mathbb{R}$ is a weight map that we introduced to give some pixels more importance in the training.

We pre-compute the weight map for each ground truth segmentation to compensate the different frequency of pixels from a certain class in the training data set, and to force the network to learn the small separation borders that we introduce between touching cells (See Figure 3c and d).

The separation border is computed using morphological operations. The weight map is then computed as

$$w(\mathbf{x}) = w_c(\mathbf{x}) + w_0 \cdot \exp\left(-\frac{(d_1(\mathbf{x}) + d_2(\mathbf{x}))^2}{2\sigma^2}\right) \tag{2}$$

where $w_c : \Omega \to \mathbb{R}$ is the weight map to balance the class frequencies, $d_1 : \Omega \to \mathbb{R}$ denotes the distance to the border of the nearest cell and $d_2 : \Omega \to \mathbb{R}$ the distance to the border of the second nearest cell. In our experiments we set $w_0 = 10$ and $\sigma \approx 5$ pixels.

In deep networks with many convolutional layers and different paths through the network, a good initialization of the weights is extremely important. Otherwise, parts of the network might give excessive activations, while other parts never contribute. Ideally the initial weights should be adapted such that each feature map in the network has approximately unit variance. For a network with our architecture (alternating convolution and ReLU layers) this can be achieved by drawing the initial weights from a Gaussian distribution with a standard deviation of $\sqrt{2/N}$, where $N$ denotes the number of incoming nodes of one neuron [5]. E.g. for a 3x3 convolution and 64 feature channels in the previous layer $N = 9 \cdot 64 = 576$.

### 3.1 Data Augmentation

Data augmentation is essential to teach the network the desired invariance and robustness properties, when only few training samples are available. In case of

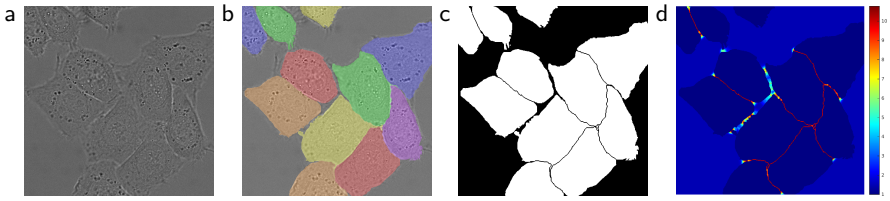


**Fig. 3.** Các tế bào HeLa trên thủy tinh được ghi lại bằng DIC (độ tương phản nhiễu di động) Mi- croscopy. **(a)** Hình ảnh thô. **(b)** Lớp phủ với phân đoạn sự thật mặt đất. Màu sắc khác nhau chỉ ra các trường hợp khác nhau của các tế bào HeLa. . .

Trong đó $\ell$: $\Omega \rightarrow \{1, \ldots, K\}$ là nhãn thực của mỗi pixel và $w$: $\Omega \rightarrow$ r là bản đồ trọng lượng mà chúng tôi giới thiệu để cung cấp cho một số pixel quan trọng hơn trong đào tạo.

Chúng tôi tổng hợp bản đồ trọng lượng cho từng phân đoạn sự thật mặt đất để kết hợp tần số khác nhau của các pixel từ một lớp nhất định trong tập dữ liệu đào tạo và buộc mạng phải tìm hiểu các đường viền phân tách nhỏ mà chúng tôi giới thiệu giữa các ô chạm (xem Hình 3C và D).

Đường viền phân tách được tính toán bằng cách sử dụng các hoạt động hình thái. Bản đồ trọng lượng sau đó được tính toán là

$$w(\mathbf{x}) = w_c(\mathbf{x}) + w_0 \cdot \exp\left(-\frac{(d_1(\mathbf{x}) + d_2(\mathbf{x}))^2}{2\sigma^2}\right) \tag{2}$$

trong đó $w_c$: $\Omega \rightarrow$ r là bản đồ trọng lượng để cân bằng các tần số lớp, $d_1$: $\Omega \rightarrow$ r biểu thị khoảng cách đến đường viền của ô gần nhất và $d_2$: $\Omega \rightarrow$ r khoảng cách đến đường viền của ô gần nhất. Trong các thử nghiệm của chúng tôi, chúng tôi đặt $w_0 = 10$ và $\sigma \approx 5$ pixel.

Trong các mạng sâu với nhiều lớp chập và các đường dẫn khác nhau qua mạng, việc khởi tạo tốt các trọng số là vô cùng quan trọng. Oth-erwise, các phần của mạng có thể kích hoạt quá mức, trong khi các phần khác không bao giờ đóng góp. Lý tưởng nhất là các trọng số ban đầu nên được điều chỉnh sao cho mỗi bản đồ tính năng trong mạng có phương sai gần đúng. Đối với một mạng với kiến trúc của chúng tôi (các lớp tích chập và relu xen kẽ), điều này có thể đạt được bằng cách vẽ các trọng số ban đầu từ phân phối Gaussian với độ lệch chuẩn của $\sqrt{2/N}$, trong đó $N$ biểu thị số lượng nút đến của một neu- ron [5]. Ví dụ. Đối với các kênh tích chập 3x3 và 64 kênh trong lớp trước $N = 9 \cdot 64 = 576$.

### 3.1 Data Augmentation

Tăng cường dữ liệu là điều cần thiết để dạy cho mạng các thuộc tính bất biến và mạnh mẽ mong muốn, khi chỉ có một vài mẫu đào tạo có sẵn. Trong trường hợp của

microscopical images we primarily need shift and rotation invariance as well as robustness to deformations and gray value variations. Especially random elastic deformations of the training samples seem to be the key concept to train a segmentation network with very few annotated images. We generate smooth deformations using random displacement vectors on a coarse 3 by 3 grid. The displacements are sampled from a Gaussian distribution with 10 pixels standard deviation. Per-pixel displacements are then computed using bicubic interpolation. Drop-out layers at the end of the contracting path perform further implicit data augmentation.

## 4 Experiments

We demonstrate the application of the u-net to three different segmentation tasks. The first task is the segmentation of neuronal structures in electron microscopic recordings. An example of the data set and our obtained segmentation is displayed in Figure 2. We provide the full result as Supplementary Material. The data set is provided by the EM segmentation challenge [14] that was started at ISBI 2012 and is still open for new contributions. The training data is a set of 30 images (512x512 pixels) from serial section transmission electron microscopy of the Drosophila first instar larva ventral nerve cord (VNC). Each image comes with a corresponding fully annotated ground truth segmentation map for cells (white) and membranes (black). The test set is publicly available, but its segmentation maps are kept secret. An evaluation can be obtained by sending the predicted membrane probability map to the organizers. The evaluation is done by thresholding the map at 10 different levels and computation of the "warping error", the "Rand error" and the "pixel error" [14].

The u-net (averaged over 7 rotated versions of the input data) achieves without any further pre- or postprocessing a warping error of 0.0003529 (the new best score, see Table 1) and a rand-error of 0.0382.

This is significantly better than the sliding-window convolutional network result by Ciresan et al. [1], whose best submission had a warping error of 0.000420 and a rand error of 0.0504. In terms of rand error the only better performing

**Table 1.** Ranking on the EM segmentation challenge [14] (march 6th, 2015), sorted by warping error.

| Rank | Group name | Warping Error | Rand Error | Pixel Error |
|---|---|---|---|---|
| | ** human values ** | 0.000005 | 0.0021 | 0.0010 |
| 1. | u-net | **0.000353** | 0.0382 | 0.0611 |
| 2. | DIVE-SCI | 0.000355 | 0.0305 | 0.0584 |
| 3. | IDSIA [1] | 0.000420 | 0.0504 | 0.0613 |
| 4. | DIVE | 0.000430 | 0.0545 | **0.0582** |
| ⋮ | | | | |
| 10. | IDSIA-SCI | 0.000653 | **0.0189** | 0.1027 |

Hình ảnh kính hiển vi mà chúng ta chủ yếu cần sự thay đổi và xoay bất biến cũng n hư độ mạnh đối với các biến dạng và biến thể giá trị xám. Đặc biệt là các biến dạn g ngẫu nhiên của các mẫu đào tạo dường như là khái niệm chính để đào tạo một m ạng phân đoạn với rất ít hình ảnh chú thích. Chúng tôi tạo ra các biến dạng trơn tru bằng cách sử dụng các vectơ chuyển vị ngẫu nhiên trên lưới 3 x 3. Các chuyển vị đ ược lấy mẫu từ phân phối Gaussian với độ lệch chuẩn 10 pixel. Các chuyển vị trên mỗi pixel sau đó được tính toán bằng cách sử dụng nội suy bicubic. Các lớp bỏ học ở cuối đường dẫn hợp đồng thực hiện tăng cường dữ liệu ngầm.

## 4 Experiments

Chúng tôi chứng minh việc áp dụng U-NET cho ba nhiệm vụ phân đoạn khác nhau. Nhiệm vụ đầu tiên là phân đoạn cấu trúc tế bào thần kinh trong các bản ghi âm điệ n tử. Một ví dụ về tập dữ liệu và phân đoạn thu được của chúng tôi được hiển thị t rong Hình 2. Chúng tôi cung cấp kết quả đầy đủ dưới dạng tài liệu bổ sung. Bộ dữ liệu được cung cấp bởi Thử thách phân đoạn EM [14] đã được bắt đầu tại ISBI 20 12 và vẫn mở cho những đóng góp mới. Dữ liệu đào tạo là một bộ gồm 30 hình ản h (512x512 pixel) từ kính hiển vi điện tử truyền qua phần nối tiếp của dây thần kin h tâm thất Instar ấu trùng của Drosophila (VNC). Mỗi hình ảnh đi kèm với một bản đồ phân đoạn sự thật được chú thích đầy đủ tương ứng cho các tế bào (màu trắng ) và màng (màu đen). Bộ thử nghiệm có sẵn công khai, nhưng bản đồ phân tích của nó được giữ bí mật. Một đánh giá có thể thu được bằng cách gửi bản đồ xác suất màng dự đoán cho ban tổ chức. Việc đánh giá được thực hiện bằng cách ngưỡng b ản đồ ở mức 10 mức và tính toán của lỗi cong vênh trên mạng, lỗi Rand Rand và lỗ i pixel pixel [14].

U-NET (tính trung bình hơn 7 phiên bản xoay của dữ liệu đầu vào) đạt được bấ t kỳ lỗi trước hoặc sau xử lý trước là 0,0003529 (điểm mới nhất, xem Bảng 1) và l ỗi RAND là 0,0382.

Điều này tốt hơn đáng kể so với kết quả mạng chập cửa sổ trượt của Ciresan et al. . Về mặt lỗi rand, hoạt động tốt hơn duy nhất

**Table 1.** Xếp hạng về Thử thách phân khúc EM [14] (ngày 6 tháng 3 năm 2015), được sắp xếp theo lỗi cong vênh.

| Rank | Group name | Warping Error | Rand Error | Pixel Error |
|---|---|---|---|---|
| | ** human values ** | 0.000005 | 0.0021 | 0.0010 |
| 1. | u-net | **0.000353** | 0.0382 | 0.0611 |
| 2. | DIVE-SCI | 0.000355 | 0.0305 | 0.0584 |
| 3. | IDSIA [1] | 0.000420 | 0.0504 | 0.0613 |
| 4. | DIVE | 0.000430 | 0.0545 | **0.0582** |
| ⋮ | | | | |
| 10. | IDSIA-SCI | 0.000653 | **0.0189** | 0.1027 |

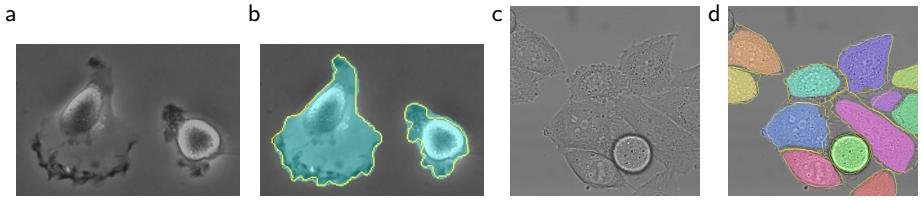


**Fig. 4.** Result on the ISBI cell tracking challenge. (**a**) part of an input image of the "PhC-U373" data set. (**b**) Segmentation result (cyan mask) with manual ground truth (yellow border) (**c**) input image of the "DIC-HeLa" data set. (**d**) Segmentation result (random colored masks) with manual ground truth (yellow border).

**Table 2.** Segmentation results (IOU) on the ISBI cell tracking challenge 2015.

| Name | PhC-U373 | DIC-HeLa |
|---|---|---|
| IMCB-SG (2014) | 0.2669 | 0.2935 |
| KTH-SE (2014) | 0.7953 | 0.4607 |
| HOUS-US (2014) | 0.5323 | - |
| second-best 2015 | 0.83 | 0.46 |
| u-net (2015) | **0.9203** | **0.7756** |

algorithms on this data set use highly data set specific post-processing methods[1] applied to the probability map of Ciresan et al. [1].

We also applied the u-net to a cell segmentation task in light microscopic images. This segmenation task is part of the ISBI cell tracking challenge 2014 and 2015 [10,13]. The first data set "PhC-U373"[2] contains Glioblastoma-astrocytoma U373 cells on a polyacrylimide substrate recorded by phase contrast microscopy (see Figure 4a,b and Supp. Material). It contains 35 partially annotated training images. Here we achieve an average IOU ("intersection over union") of 92%, which is significantly better than the second best algorithm with 83% (see Table 2). The second data set "DIC-HeLa"[3] are HeLa cells on a flat glass recorded by differential interference contrast (DIC) microscopy (see Figure 3, Figure 4c,d and Supp. Material). It contains 20 partially annotated training images. Here we achieve an average IOU of 77.5% which is significantly better than the second best algorithm with 46%.

## 5 Conclusion

The u-net architecture achieves very good performance on very different biomedical segmentation applications. Thanks to data augmentation with elastic defor-

[1] The authors of this algorithm have submitted 78 different solutions to achieve this result.

[2] Data set provided by Dr. Sanjay Kumar. Department of Bioengineering University of California at Berkeley. Berkeley CA (USA)

[3] Data set provided by Dr. Gert van Cappellen Erasmus Medical Center. Rotterdam. The Netherlands

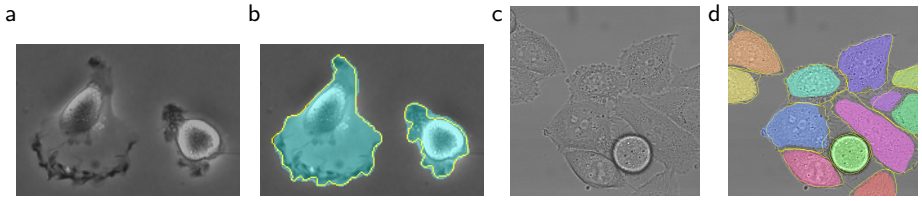


**Fig. 4.** Kết quả trên Thử thách theo dõi ô ISBI. . . .

**Table 2.** Kết quả phân đoạn (IOU) trên Thử thách theo dõi ô ISBI 2015.

| Name | PhC-U373 | DIC-HeLa |
|---|---|---|
| IMCB-SG (2014) | 0.2669 | 0.2935 |
| KTH-SE (2014) | 0.7953 | 0.4607 |
| HOUS-US (2014) | 0.5323 | - |
| second-best 2015 | 0.83 | 0.46 |
| u-net (2015) | **0.9203** | **0.7756** |

Các thuật toán trên tập dữ liệu này Sử dụng các phương thức xử lý hậu xử rất cao về các phương pháp xử lý sau [1] được áp dụng cho bản đồ xác suất của Ciresan et a l. [1].

Chúng tôi cũng đã áp dụng NET U vào một nhiệm vụ phân đoạn tế bào trong cá c thời gian hiển vi bằng kính hiển vi ánh sáng. Nhiệm vụ Segmenation này là một ph ần của Thử thách theo dõi ô ISBI 2014 và 2015 [10,13]. Bộ dữ liệu đầu tiên của Phc Phc-U373, [2] chứa các tế bào U373 của Glioblastoma-astrocytoma U373 trên chất nề n polyacrylimide được ghi lại bằng kính hiển vi tương phản pha (xem Hình 4A, B v à Supp. Vật liệu). Nó chứa 35 hình ảnh đào tạo được chú thích một phần. Ở đây, c húng tôi đạt được mức IOU trung bình (giao lộ trên Union Union) là 92%, tốt hơn đ áng kể so với thuật toán tốt thứ hai với 83% (xem Ta- ble 2). Bộ dữ liệu thứ hai tập hợp DIC DIC-HELA [3] là các tế bào HeLa trên một kính được ghi lại bằng kính hiển vi tương phản nhiễu (DIC) (xem Hình 3, Hình 4C, D và Supp. Vật liệu). Nó chứa 20 hình ảnh đào tạo chú thích một phần. Ở đây chúng tôi đạt được mức trung bình IO U là 77,5%, tốt hơn đáng kể so với thuật toán tốt thứ hai với 46%.

## 5 Conclusion

Kiến trúc U-NET đạt được hiệu suất rất tốt trên các ứng dụng phân đoạn y học rất khác nhau. Nhờ gia tăng dữ liệu với độ đàn hồi-

1 Các tác giả của thuật toán này đã gửi 78 giải pháp khác nhau để đạt được kết quả này. [2] Bộ dữ liệu được cung cấp bởi Tiến sĩ Sanjay Kumar. Khoa Đại học Sinh học California tại Berkeley. Berkeley CA (Hoa Kỳ) [3] Bộ dữ liệu được cung cấp bởi Trung tâm y tế bác sĩ Gert van Cappellen Erasmus. Rotterdam. Hà Lan

mations, it only needs very few annotated images and has a very reasonable training time of only 10 hours on a NVidia Titan GPU (6 GB). We provide the full Caffe[6]-based implementation and the trained networks[4]. We are sure that the u-net architecture can be applied easily to many more tasks.

## Acknowlegements

This study was supported by the Excellence Initiative of the German Federal and State governments (EXC 294) and by the BMBF (Fkz 0316185B).

## References

1. Ciresan, D.C., Gambardella, L.M., Giusti, A., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. In: NIPS. pp. 2852–2860 (2012)
2. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: NIPS (2014)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
4. Hariharan, B., Arbelez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization (2014), arXiv:1411.5752 [cs.CV]
5. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification (2015), arXiv:1502.01852 [cs.CV]
6. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding (2014), arXiv:1408.5093 [cs.CV]
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1106–1114 (2012)
8. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Computation 1(4), 541–551 (1989)
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation (2014), arXiv:1411.4038 [cs.CV]
10. Maska, M., (...), de Solorzano, C.O.: A benchmark for comparison of cell tracking algorithms. Bioinformatics 30, 1609–1617 (2014)
11. Seyedhosseini, M., Sajjadi, M., Tasdizen, T.: Image segmentation with cascaded hierarchical models and logistic disjunctive normal networks. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 2168–2175 (2013)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014), arXiv:1409.1556 [cs.CV]
13. WWW: Web page of the cell tracking challenge, http://www.codesolorzano.com/celltrackingchallenge/Cell_Tracking_Challenge/Welcome.html
14. WWW: Web page of the em segmentation challenge, http://brainiac2.mit.edu/isbi_challenge/

[4] U-net implementation, trained networks and supplementary material available at http://lmb.informatik.uni-freiburg.de/people/ronneber/u-net

Mations, nó chỉ cần rất ít hình ảnh chú thích và có thời gian đào tạo rất hợp lý chỉ 10 giờ trên GPU NVIDIA Titan (6 GB). Chúng tôi cung cấp thực hiện dựa trên Ca ff e đầy đủ [6] và các mạng được đào tạo [4]. Chúng tôi chắc chắn rằng kiến trúc U-NET có thể được áp dụng dễ dàng cho nhiều nhiệm vụ hơn.

## Acknowlegements

Nghiên cứu này được hỗ trợ bởi Sáng kiến Xuất sắc của Chính phủ Liên bang và Ti ểu bang Đức (Exc 294) và BMBF (FKZ 0316185B).

## References

1. Trong: NIP. Trang 2852 Từ2860 (2012) 2. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T. Trong: NIPS (2014) 3. Girshick, R., Donahue, J., Darrell, T., Malik, J .: Phân cấp tính năng phong phú để phát hiện đối tượng và phân đoạn ngữ nghĩa. Trong: Thủ tục tố tụ ng của Hội nghị IEEE về Tầm nhìn máy tính và nhận dạng mẫu (CVPR) (2014) 4. Hariharan , B., Arbelez, P., Girshick, R., Malik, J. J .: Xuất sâu vào trực tràng: Vượt qua hiệu suất cấp độ con người trên ImageNet Classi fi cation (2015), Arxiv: 1502.01852 [CS.CV] 6. Jia, Y., S helhamer, E., Donahue, J. Nhúng tính năng nhanh (2014), ARXIV: 1408.5093 [CS.CV] 7. Kri zhevsky, A., Sutskever, I., Hinton, G.E .: Phân loại ImageNet với mạng lưới thần kinh liên tụ c sâu. Trong: NIP. Trang 1106 Từ1114 (2012) 8. Tính toán thần kinh 1 (4), 541 Từ551 (1989) 9

10. Maska, M., (...), De Solorzano, C.O .: Một chuẩn mực để so sánh các thuật toán theo dõi t ế bào. Tin sinh học 30, 1609 Từ1617 (2014)
11. Seyedhosseini, M., Sajjadi, M., Tasdizen, T .: Phân đoạn hình ảnh với các mô hình phân cấ p xếp tầng và các mạng bình thường khác nhau. Trong: Tầm nhìn máy tính (ICCV), Hội nghị quốc tế IEEE 2013 về. trang 2168 Từ2175 (2013)
12. Simonyan, K., Zisserman, A .: Mạng tích chập rất sâu để nhận dạng hình ảnh quy mô lớn (2014), ARXIV: 1409.1556 [CS.CV]
13. WWW: Trang web của Thử thách theo dõi ô, `http://www.codesolorzano.com/celltrackingchallenge/Cell_Tracking_Challenge/Welcome.html`
14.

[4] U-net implementation, trained networks and supplementary material available at http://lmb.informatik.uni-freiburg.de/people/ronneber/u-net