A DEEP LEARNING METHOD WITH CRF FOR INSTANCE SEGMENTATION OF METAL-ORGANIC FRAMEWORKS IN SCANNING ELECTRON MICROSCOPY IMAGES

Ilyes Batatia

Laboratoire PMC, Ecole Polytechnique-CNRS, IP Paris, 91128 Palaiseau, France
University Paris Saclay, ENS Paris-Saclay, 94230 Cachan, France (ilyes.batatia@ens-paris-saclay.fr)

ABSTRACT

This paper proposes an integrated method for recognizing special crystals, called metal-organic frameworks (MOF), in scanning electron microscopy images (SEM). The proposed approach combines two deep learning networks and a dense conditional random field (CRF) to perform image segmentation. A modified Unet-like convolutional neural network (CNN), incorporating dilatation techniques using atrous convolution, is designed to segment cluttered objects in the SEM image. The dense CRF is tailored to enhance object boundaries and recover small objects. The unary energy of the CRF is obtained from the CNN. And the pairwise energy is estimated using mean field approximation. The resulting segmented regions are fed to a fully connected CNN that performs instance recognition. The method has been trained on a dataset of 500 images with 3200 objects from 3 classes. Testing achieves an overall accuracy of 95.7% MOF recognition. The proposed method opens up the possibility for developing automated chemical process monitoring that allows researchers to optimize conditions of MOF synthesis.

Index Terms— metal-organic frameworks, semantic segmentation, deep learning, conditional random fields

1. INTRODUCTION

Metal-Organic Frameworks (MOF) are special type of crystals that attract an important research interest in various scientific and engineering domains [1], such as gas storage [2], sensing [3], catalysis [4] or separation [5]. The experimental conditions of their growth lead to a variety of sizes and geometries. Engineering MOFs consists mainly in controlling such conditions for selectively forming the crystal with the geometry and the size of interest. Imaging is an important tool for this control process. Various techniques are used for characterizing MOFs, such as atomic force microscopy (AFM) which provides topographic information, and X-Ray Diffraction for internal structure. However, scanning electron microscope (SEM) remains an important source of information for this phenomena. SEM is an imaging technique designed to detect and quantify secondary electrons backscattered from a material surface, using an Everhart-Thornley detector. It forms an image that characterizes the surface of the material with a high resolution, typically a few nanometers. This imaging modality has a wide range of applications. It is used since a long time for the analysis of crystal phenomena [6, 7]. It is also commonly used to monitor visually MOF synthesis processes [8]. Various SEM image segmentation methods have been reported in the literature [9], with application to many domains. However, there are no reports of specific work on automated SEM image segmentation with application to MOF. Such images exhibit a complex nature with clutter and occlusion of multiple types of particles with different 3D orientation, shape and size (Fig.1).

Currently, researchers analyse manually SEM images, using simple processing tools. Automatic image processing methods are required to assist in this fastidious and time-consuming task. They are also needed to provide means to control the synthesis conditions in order to selectively favour MOF with preferred geometry. This paper, contributes to this effort by proposing an automatic semantic segmentation method to identify, localize, and recognize particles in SEM images and quantify their abundance. The proposed method relies on spatially coherent deep learning to segment and recognize particles despite clutter and occlusion.

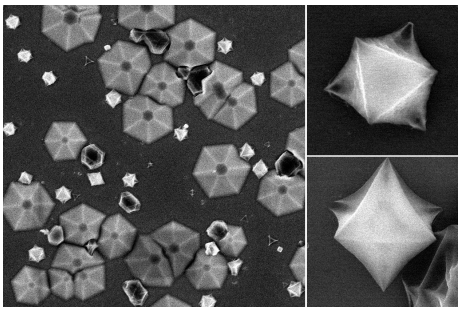

Fig. 1. Example of SEM images of MOFs

Semantic segmentation is a class of algorithms that aim at extracting information from images and using such information to accomplish some specific tasks. The traditional approach to semantic segmentation was based on conditional random fields (CRF) to correlate pixels using probabilistic models [10]. The large development of deep learning during the last years [11] led to the reformulation of many machine learning tasks, including segmentation [12], and semantic segmentation [13]. Most existing semantic segmentation methods incorporate CRF, either as post-processing, as regularization of the loss function, or as a layer into the deep network. Despite the excellent results obtained by many of these methods, the true instance-segmentation task is still a challenge [14].

This paper proposes a cascade of deep networks that perform MOF instance segmentation in SEM images. The first contribution of this paper is an original deep learning architecture customized for recognizing a variety of objects despite their clutter, occlusion and similarity. Precisely, the architecture consists of a convolutional network that segments objects. This Unet-like network incorporates dilatation layers with atrous convolutions, allowing to consider the image context. The resulting segmentation is fed to a conditional random field that refines the segmentation by enforcing the spatial coherence of objects. Objects are then automatically extracted and fed individually to a third network that classifies each object. The second contribution is related to the application domain. To the best of our knowledge, this is the first work that proposes an image based automatic quantitative method to analyse MOF objects. We show experimentally that the proposed architecture is suitable for our ap-

plication due to the existence of small objects that other methods ignore. Its second advantage lies in its ability to require a fairly small number of images, compared to other deep learning methods. This is particularly critical for thin-film MOF applications, where chemical experimental conditions greatly complicate the creation of large image datasets. Nevertheless, the obtained results are very promising and show the potential of the proposed method to develop useful quantitative tools for computational chemistry in the MOF domain.

The remainder of the paper is structured as follows. The proposed method is presented in Section 2. The overall architecture of our solution is first presented. Each component of this architecture is then detailed in the subsequent subsections. Section 3 describes experimental results using real images obtained in the laboratory. Quantitative evaluation of the results of the different parts of our architecture are presented, compared, and discussed. Finally, conclusions and future work are reported in Section 4.

2. PROPOSED METHOD

The proposed method composes of four components (Fig.2). The SEM image is given as input to a segmentation deep convolutional neural network, called dilated-unet, that produces a segmentation mask. The segmentation result has often many flaws, especially with regards to small objects and boundary precision. To remedy such shortcomings, the segmentation is improved using a conditional random field that enforces the spatial coherence of objects and recovers small objects. The third component of the proposed cascade architecture consists of extracting separate objects (or regions). Thumbnail images of the objects are extracted from the original image and passed to a second CNN that performs object classification. These four components are explained in the following sections.

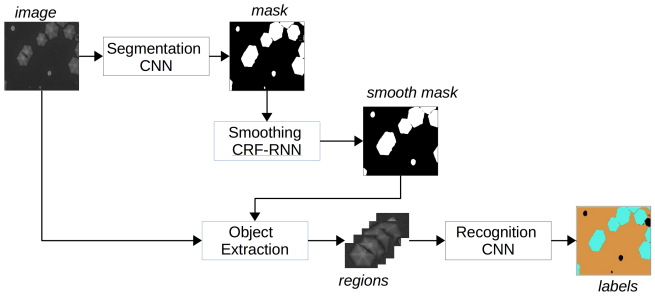

Fig. 2. Block diagram showing the main components of the proposed method

2.1. Dilated-Unet for segmentation

In order to segment objects within a 512×512 SEM images, we designed a Unet-like convolutional neural network, denoted *dilated-Unet*. As depicted in figure (Fig.3), the down branch of the network consists of a pattern of two convolutions followed by 2×2 max-pooling. We denote the pattern by (S_{c1}, S_{c2}, D_c), where S_{c1}, S_{c2} are respectively the number of output feature maps of the first and second convolution, and D_c is the size of the output feature maps of the convolutions. Obviously, the max-pool layer will divide this size by 2 before feeding it to the subsequent layer. This pattern is repeated four times, with output numbers and sizes $\{(64, 64, 512), (128, 128, 256), (256, 256, 128), (512, 512, 64)\}$, respectively. This is followed by the bottleneck phase made of two convolutions of 1024 outputs feature maps with size 32×32.
The main difference with the classical Unet architecture consists in introducing a chain of atrous convolutions. It has been argued in [15], that such algorithm allows one to compute the responses of any layer at any desirable resolution. The technique consists of performing a sort of dilatation of the convolution filter by inserting zero lines and columns. The number of such inserted lines/columns is usually called the dilatation rate. Although such filter has increased size, by implementing it through sparse representation, the number of filter parameters and of convolution operation remain unchanged, compared to the original filter. We introduced in our network a chain of three atrous convolution layers (green layers in Fig. 3), with rates 4, 16 and 32, respectively. This change in architecture justifies the name dilated-unet. These operations allow us to significantly increase the resolution of the output features. This is motivated by the need to consider favourably the very high resolution of SEM images. The up branch of the network consists of a pattern of up-sampling followed by two up-convolutions. This pattern is repeated four times to recover the image original resolution. Finally, a soft-max layer estimates the probabilities of pixels belonging to objects or to the background. The zero/one label map is then outputted. The network parameters were estimated at training stage using a stochastic gradient descent to minimize the cross entropy loss function.

2.2. CRF for boundary correction

The segmentation mask produced by our dilated-unet exhibits irregularities around objects and under-estimate small objects. Various methods have been reported in the literature to remedy this problem [15]. Among these approaches, fully connected conditional random field (CRF) has shown good results in combining the detection ability of CNN and precision of boundaries given by CRF. A conditional random field is a probabilistic graph $\mathcal{G}(\boldsymbol{y}, \mathcal{E})$ defined by its nodes $\boldsymbol{y}$ and edges $\mathcal{E}$. Such graph is modelled by the Gibbs distribution $p(\boldsymbol{y}) = (1/Z)\exp(-E(\boldsymbol{y}))$, where Z is a normalizing constant and $E(\boldsymbol{y})$ an energy. In this work, the second component of our cascade implements the *Dense CRF* method proposed in [16].
Without loss of generality consider a two class CRF. Let $\mathcal{L} = \{0, 1\}$ the set of labels, $\{x_1, ..x_N\}$ the set of random variables corresponding to individual image pixels, and $\{y_1, ..y_N\}$ latent variables representing the nodes of the CRF, defined so that y_i^ℓ indicates whether pixel i has label $\ell \in \mathcal{L}$. The energy associated with label y is given by the following general expression

$$E(\boldsymbol{y}) = \sum_i \psi_u(y_i^\ell) + \sum_{(i,j)\in\mathcal{E}} \psi_p(y_i^\ell, y_j^m) \qquad (1)$$

where $\boldsymbol{y} \in \mathcal{L}$ is the set of latent variables, and $\mathcal{E}$ is the set of edges of the CRF representing relation between pairs of pixels. In dense CRF, $\mathcal{E} = \{(i, j), \forall i \neq j\}$, the sum is the second term can be done for $i \neq j$. The unary potential term $\psi_u(y_i^\ell)$ is calculated based on the segmentation given by our dilated-unet. Precisely, let $p(y_i^{(1)})$ the probability, given by the soft-max layer, that pixel i belongs to an object (not to the background, whose class is zero). The unary potential is given by $\psi_u(y_i^{(1)}) = -\log p(y_i^{(1)})$. The pairwise potential $\psi_p(y_i^\ell, y_j^m)$ penalizes two pixels i and j assigned to classes ℓ and m. Akin to [16], we adopt a pairwise potential consisting of a mixture of two Gaussian kernels:

$$\psi_p(y_i^\ell, y_j^m) = \mu(\ell, m)\left[\omega_1 \exp(-|x_i - x_j|^2) + \omega_2 \exp -(||p_i - p_j||^2)\right]$$

where x_i and x_j are pixel values, p_i and p_j are coordinates of pixels i and j, and $||.||^2$ is the norm of a vector. The weight hyperparameters ω_1 and ω_2 are estimated by cross-validation. The term $\mu(\ell, m)$, called the label compatibility function, expresses prior knowledge

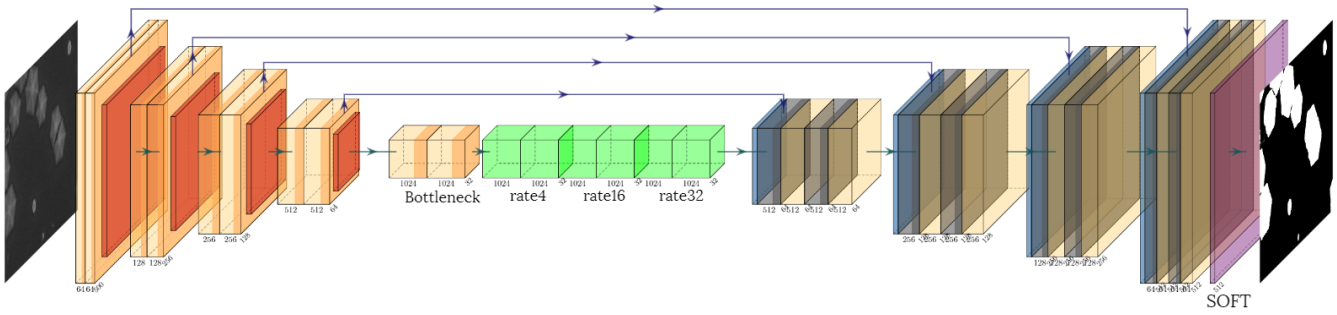

Fig. 3. Unet-like CNN with atrous convolution for image segmentation. Dimensions are detailed in the text (Section 2.1).

on pairs of labels (ℓ, m) being adjacent. In our case, no specific prior knowledge exists, therefore we adopt the Potts model where $\mu(\ell, m) = \delta_{\ell m}^{-1}$, with δ the Kronecker delta. Performing inference using such CRF consists in estimating the distribution $p(\boldsymbol{y})$. This task is intractable and approximations are proposed in the literature. The mean field approximation has been shown to be appropriate for the dense CRF [16]. We implemented this algorithm as a recurrent neural network using fast filtering technique from [17]. The result is an improved segmentation mask where boundary are regularized and small objects better recovered.

2.3. Object extraction

This component of our cascade method (Fig.2) consists in extracting a thumbnail image centered on each object present in the scene for subsequent classification. In the absence of occlusion and overlap of objects, the segmentation mask would provide entirely isolated and connected region for each object. Extraction of such isolated regions in black and white images is straightforward, using for example a connected component algorithm. However, in our case, SEM images are cluttered and objects often have common borders, due to proximity that cannot be resolved at the image resolution. Therefore, despite the boundary correction of the CRF, objects must be separated before determining their bounding boxes. For this purpose, we implemented the Chan and Vese level set algorithm [18] as complementary variational segmentation method. The level set is initialized with the result of the CRF, providing therefore a very good initialization, that improves convergence and speed. By definition, the level set minimizes the variance within objects. This makes it appropriate to detect fine details due to intensity variations. This property combined to the good initialization allows a better separation of objects. Following separation, connected components are determined and bounding boxes are calculated. The corresponding image thumbnails are extracted. Please note that thumbnails are image rectangular regions centered on the corresponding objects. However, in addition to the considered object, a thumbnail often contains parts of the neighbouring objects. This choice has been made to allow the network to learn neighbourhood rules. Thumbnails are fed to the next component of our method for classification of the object at its center, as explained in the next section.

2.4. Object recognition

The size-normalized (224×224) thumbnail images extracted after segmentation are used as input to a classification CNN having the standard VGG16 architecture (Fig. 4). This network was trained to classify each image according to the type of object at its center. It is important to note that a thumbnail image contains one object at its center, but also parts of neighbouring objects. The central object can be of any size, as our chemical phenomena can grow objects of similar types but with varied sizes. The classification CNN has been trained to learn object types despite the variability of size and type of neighbours.

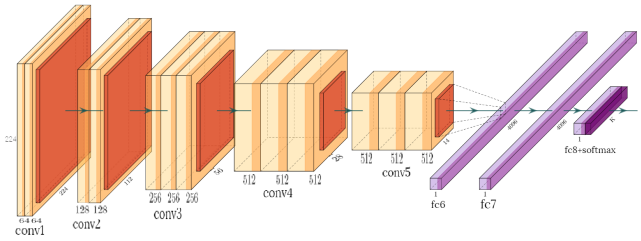

Fig. 4. Architecture of the image classification

3. EXPERIMENTS

3.1. Dataset

We used $Fe3+$ and benzene $1,4$ dicarboxylic acid in DMF at a temperature of $70°C$ to $120°C$ to grow MOFs of interest, on a surface of $Si[111]$ plane. The grown MOFs were imaged using scanning electron microscopy (SEM). Raw 766×1088 images were denoised prior to processing. We had three types of objects: {Hexagon, Hexagon 2, Octahedron}. Ground truth labeling was performed manually on a set of 100 images containing 3200 individual objects, each assigned a different color. Given the low number of images, overfitting can occur. To reduce this risk, a data augmentation strategy was applied. Multiple 512×512 images are extracted for each 766×1088 images using different cropping methods. In addition, the original images were resized to 512×512. This resulted in about 552 images of $512x512$ pixels. The dataset was divided into two sets: 500 images for training and 52 images for testing. Rotations and scaling were applied to the first 500 images, obtaining a training dataset of total 64000 images of 512×512.
In addition, for training the object classification CNN (fourth component of our method, Section.2.4), each crystal object was extracted individually in a separate thumbnail image. The size of all these thumbnails was normalized to 224×224, using the *image data augmentation* toolbox under matlab®. A set made of 1000 images from each object type was used for training, giving 3000 images, with balanced classes. These images underwent rotations and scaling giving a total of 180000 images of size 224×224 of individual objects. An extra set of 100 images was made for the validation of the classification CNN.

3.2. Results and comparison

Segmentation results: The dilated U-net was trained with 64000 images of 512×512, using a GPU Nvidia GTX 1660. A stochastic gradient descent optimization algorithm was used, with 40 epochs and 4 batch size. In order to avoid overfitting, accuracy and loss function curves were monitored at the training stage (Fig.5). The learning rate hyper-parameter was consequently set to 10^{-4}. The final accuracy was 99.15% over the 8 epochs. Figure 6 shows an example of segmented image. One notices visually, a good segmentation for most objects. In order to assess the quality of the segmentation, we calculated the dice function and obtained the average score of 0.8487 over the entire test dataset.

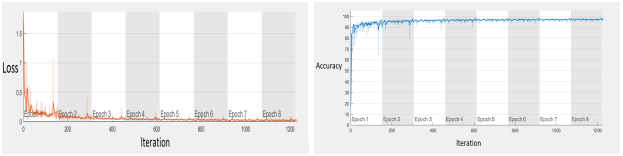

Fig. 5. Curves of the loss function and the accuracy during the training stage of the dilated-Unet.

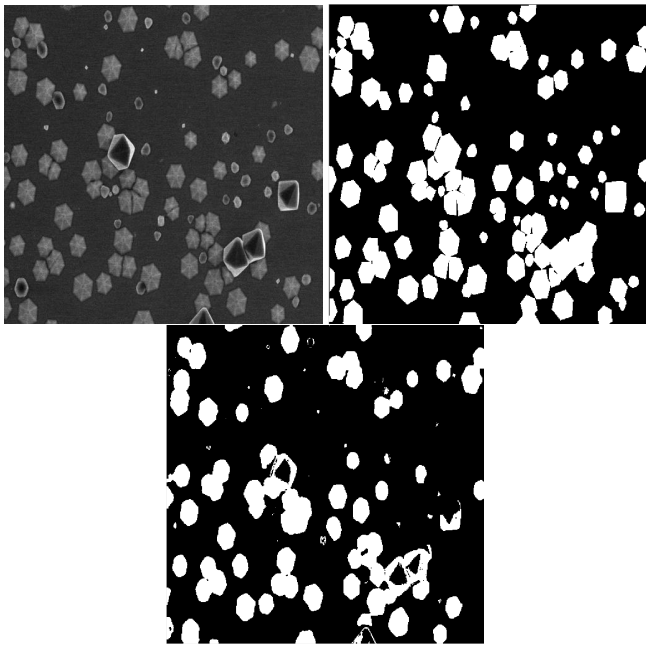

Fig. 6. Example of segmentation result, with the original image (top left), the ground truth (top right), and the segmentation result (bottom)

CRF results: Segmentation results were given as input to the CRF layer. The unary potential was extracted from the softmax output map. The mean field algorithm was run 20 iterations to estimate the random field distribution. The process of segmentation by the dilated-unet followed by the CRF was applied to the 52 images of our test dataset (Section 3.1). Figure 7 shows an example of segmentation refinement by the CRF. One notices the recovery of a number of small objects and the correction of the objects boundaries. The dice was calculated on these results, giving an average score of 0.8867. Compared to 0.8487 from the dilated-unet, one notes the increase by 3.8% of the overall surface of segmented regions. Although this global measure seems modest, the local effect is important as mentioned above.

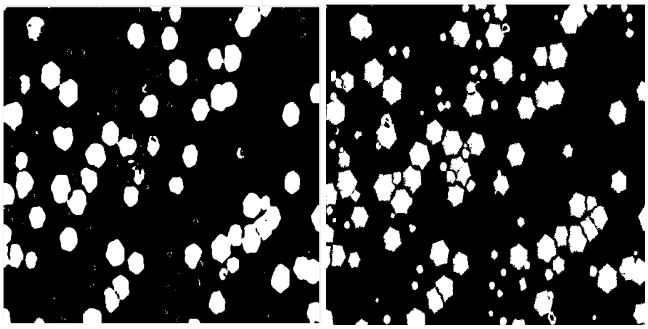

Fig. 7. Example of segmentation using the dilated-unet (left) and corresponding CRF refinement (right).

Classification results: The 224×224 image thumbnails extracted based on the objects bounding boxes were given to the classification CNN. After training, we tested the network on 52 randomly selected images from the test dataset, containing a total of 1358 objects, with $(369, 186, 803)$ from Hexagon, Hexagon 2, Octahedron, respectively. In order to assess the classification performance, we computed the confusion matrix (Table 1). The overall accuracy was 95.7%. The sensitivity, precision, specificity and accuracy for each class were then computed (Table 2). One notices the good performance, except for Hexagon2 which is a weak class (186 instances out of 1358).

Table 1. Confusion matrix of the classification CNN

	Hexagon	Hexagon 2	Octahedron
Hexagon	342	25	2
Hexagon 2	5	161	20
Octahedron	0	10	793

Table 2. Performance metrics of the classification CNN

	Sensitivity	Precision	Specificity	Accuracy
Hexagon	92.7%	98.6%	99.5%	97.6%
Hexagon 2	86.6%	82.1%	97.0%	95.6%
Octahedron	98.7%	97.3%	96.0%	97.6%

Comparison: In order to compare our method, we implemented a pixel-wise semantic segmentation method from the literature, named Resnet18 [19]. Figure 8 shows example of compared results. Our method outperforms the other method, especially in terms of detecting and recognizing small objects, which belong to weak class. Adding techniques such as weighted loss function to Resnet18 did not solve the problem. This illustrates the two advantages of our method (handling clutter and small objects).

4. CONCLUSIONS

We presented a deep learning cascade method for recognizing MOF crystals in SEM images from chemical experiments. The method consists of an enhanced Unet-like CNN called *dilated-Unet* that includes atrous convolutions to consider the context of the image for enhancing the resolution of feature maps. Segmentation obtained by this network is improved using a *dense-CRF* implemented using mean field approximation. This graphical model is tailored to correct boundaries and recover small objects. Thumbnail images

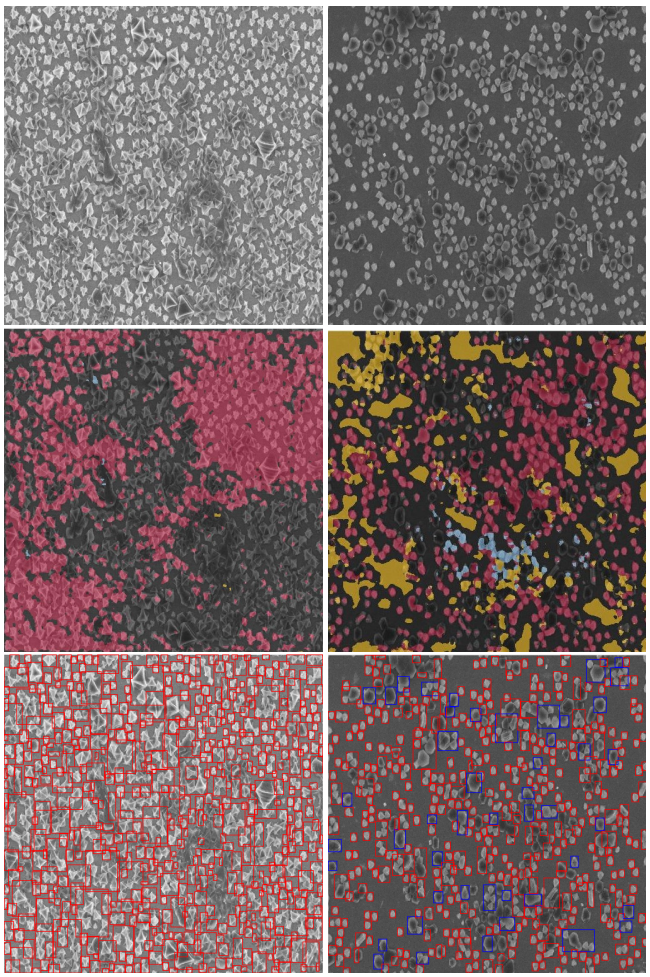

Fig. 8. Example of segmentation result, with the original images (top), semantic segmentation [19] (middle), and our method (bottom). For the first image, the semantic method misses a number of small and average objects, which are recognized by our method. For the second, our method detects the two types of objects contrarily to the semantic method, which mixes objects.

centered on objects are then extracted and classified using a classification CNN to recognize objects. Experimental results with a moderate dataset (due to difficulty of obtaining the MOF phenomena) showed very good results.

Future work will focus on better methods to separate objects (Section 2.3), using deep learning and shape priors. In addition, we started working on deep learning for in-object segmentation to identify facets, at the scale of thumbnails. Results will allow us to measure metrics on objects such as orientation, growth rate.

5. ACKNOWLEDGEMENT

I would like to express my gratitude to Catherine Henry-de-Villeneuve and Michel Rosso for welcoming me into their team at Ecole Polytechnique. I also thank them and Weichu Fu for providing me with raw images and detailed information on the underlying chemical phenomena and for many fruitful discussions about my results.

6. REFERENCES

[1] H. Furukawa, K. E. Cordova, M. O'Keeffe, and O. M. Yaghi, "The chemistry and applications of metal-organic frameworks," *Science*, vol. 341, no. 6149, p. 1230444, 2013.

[2] K. Sumida, D. L. Rogow, J. A. Mason, T. M. McDonald, E. D. Bloch, Z. R. Herm, T.-H. Bae, and J. R. Long, "Carbon dioxide capture in metal–organic frameworks," *Chemical reviews*, vol. 112, no. 2, pp. 724–781, 2012.

[3] H. Li, X. Feng, Y. Guo, D. Chen, R. Li, X. Ren, X. Jiang, Y. Dong, and B. Wang, "A malonitrile-functionalized metal-organic framework for hydrogen sulfide detection and selective amino acid molecular recognition," *Scientific reports*, vol. 4, p. 4366, 2014.

[4] B. An, J. Zhang, K. Cheng, P. Ji, C. Wang, and W. Lin, "Confinement of ultrasmall cu/zno x nanoparticles in metal–organic frameworks for selective methanol synthesis from catalytic hydrogenation of co2," *Journal of the American Chemical Society*, vol. 139, no. 10, pp. 3834–3840, 2017.

[5] S. Qiu, M. Xue, and G. Zhu, "Metal–organic framework membranes: from synthesis to separation application," *Chemical Society Reviews*, vol. 43, no. 16, pp. 6116–6140, 2014.

[6] G. E. Lloyd, "Atomic number and crystallographic contrast images with the sem: a review of backscattered electron techniques," *Mineralogical Magazine*, vol. 51, no. 359, pp. 3–19, 1987.

[7] H. Dong and G. M. Koenig, "A review on synthesis and engineering of crystal precursors produced via coprecipitation for multicomponent lithium-ion battery cathode materials," *CrystEngComm*, 2020.

[8] O. Shekhah, J. Liu, R. Fischer, and C. Wöll, "Mof thin films: existing and future applications," *Chemical Society Reviews*, vol. 40, no. 2, pp. 1081–1106, 2011.

[9] M. Salzer, T. Prill, A. Spettl, D. Jeulin, K. Schladitz, and V. Schmidt, "Quantitative comparison of segmentation algorithms for fib-sem images of porous media," *Journal of microscopy*, vol. 257, no. 1, pp. 23–30, 2015.

[10] A. Arnab, S. Zheng, S. Jayasumana, B. Romera-Paredes, M. Larsson, A. Kirillov, B. Savchynskyy, C. Rother, F. Kahl, and P. H. Torr, "Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 37–52, 2018.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[13] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41–65, 2018.

[14] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3150–3158.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[16] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[17] A. Adams, J. Baek, and M. A. Davis, "Fast high-dimensional filtering using the permutohedral lattice," in *Computer Graphics Forum*, vol. 29, no. 2. Wiley Online Library, 2010, pp. 753–762.

[18] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001.

[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.