Multi-focus image fusion with a deep convolutional neural network for Semiconductor Inspection tools

Srinivasa Rao (srao78@stanford.edu), Antariksh De (antde@stanford.edu)
GitHub Project Link: https://github.com/antarikshde/DeepLearning_ImageFusion

***Abstract*—** Multi-Focus Image fusion (MFIF) is an important technique to reconstruct a fully focused image (FFI) from two or more partly focused images of the same scene. In semiconductor industry, as chip design gets more complicated, capturing FFI gets harder due to topological differences on a wafer (varying heights of the structures). Traditional Computer Vision techniques take multiple source images from the same location at different focal offsets to generate a useful FFI for inspection tools and is time consuming. We propose a deep supervised model for the generation of FFI to solve semiconductor inspection image defocus issue in less time in order to increase productivity throughput of these tools. After about 50 epochs, our network converged to a decent final focused image which can be used for semiconductor wafer inspection. Our test set gave us 98% good results.

I. Introduction

In Semiconductor inspection tools, it is difficult for inspection/review cameras to take images from top with all 3D structures on a wafer in focus since the structures are at different focal plane. For a given focal setting, only the structures within the depth-of-field (DOF), appear to be sharp in the image whereas the other structures remain defocused. In recent years, many image fusion techniques have been proposed which are classified into 2 categories: transform domain and spatial domain [6]. These techniques require a lot of input images. In this paper, we address this problem with a deep learning approach by learning a mapping between multiple blurry images of a wafer site to a fully focused image.

To the best of our knowledge, this is the first time, Deep Learning (DL) is used for image fusion in semiconductor domain. The novelty is using DL for image fusion using a convolution neural network (CNN) for a totally new application. In recent years, CNNs have been used for visible-infrared image fusion, medical image fusion and multi-exposure image fusion for but not in semiconductor space. For any semiconductor inspection tool, throughput is a major concern and thus our proposed network solves this problem with minimum input images for a real-time solution with less runtime computation complexity. During training, our network expects multiple (1 to n) input images of any size and outputs a fully focused image of size 256x256 whereas during inference, the network expects multiple square images of any size and outputs a fully focused image of the same size as the input.

This paper describes other image fusion Related-Work in Section2. Section 3 talks about the Dataset including training, dev and test set. The Model/Network is described in Section 4, the Results/Experiments in Section 5, Conclusion/Future-Work in Section 6 and Contributions in Section 7.

II. Related Work

Advantages of a simple pixel-based image fusion (which averages the pixel values) are simple and fast, but it tends to blur the image losing some of its information. Several state-of-the-art pixel-based image fusion algorithms have been proposed, such as guided filtering [7] and dense SIFT [8] to overcome the above mentioned pitfalls. Guided filtering and dense SIFT first generate the fusion map by detecting the focused pixels from each source image; then, based on the modified decision map, the final fused image is obtained by selecting the pixels in the focus areas.

Of late, there has been some work related to image fusion using Deep Learning. Liu et al. [1] proposed a deep network and used popular image databases for their training data and added Gaussian blur to simulate multi-focus images. They classified their images into focused and unfocused pixels and generated an initial focus map. They performed some post processing in order to get the final fully focused image. A lot of compute power is needed for this network and it can only work for bi-modal blurred images. Tang et.al [2] worked on multi-focus image fusion. The authors also generated defocused images by automatically adding blur to the original images. The output of the model are three probabilities: defocused, focused or unknown for each pixel. This network also needed post processing and is compute intensive. Our model can take more than 2 input images without being compute intensive and yet be fast and simple.

III. Dataset And Features

The data set was collected using multiple customer wafers at multiple focal offsets and at various high topology feature sites, using KLA-Tencor's proprietary high resolution semiconductor wafer inspection system.

- Collected 20 images at different focal offsets at 100 different sites from the wafer, for a total of 2000 images.
- This paper uses the 90/5/5 approach for train/dev/test set respectively.
- From the 100 different sites, 90 sites (1800 images) were used for training, 5 sites (100 images) as dev set, and 5 sites (100 images) for test/validation.
- Each image is a colored image of size 640 x 480 pixels but gets converted to 256 x 256.
- The maximum topology difference at a given site was about 20 microns.

Fully Focused images (Ground truth) were generated using KLA-Tencor's proprietary Software for each high topology site. For our project, the input images fed to the network are raw color images taken from different sites on a semiconductor wafer. No pre-processing was required on the input images as it can compromise the final image quality. Image quality is of utmost importance for semiconductor inspection systems. Examples of input images are shown below (Fig. 1.)

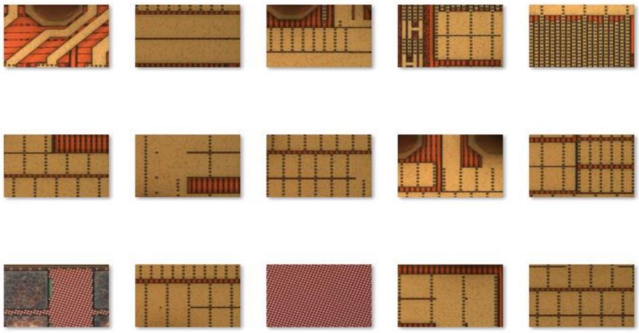


Figure. 1. Examples of Input data set.

IV. Methods

CNNs have several convolution/pooling layers followed by fully connected layers. The first part is viewed as feature extractors while the later as classifiers. Since Image fusion is a combination of the two: high frequency detail extraction and clarity information classification to classify different source images, CNNs can be feasible for Image fusion. Usually in CNN based image fusion methods, only the last layer results are used as image features and loses most of the important information from the middle layers. This paper uses a feature extraction and image reconstruction architecture together with image fusion. The schematic diagram of our architecture is shown in Fig. 2. The feature extraction consists of convolution layers and dense blocks in which the output of each layer is used as the input of next layer. The image reconstruction includes four CNN layers. For our simplicity, the source images are two blurred images instead of multiple images even though this architecture can be extended for multi-defocused images as well.

A. Model

As shown in Table I. the feature extractor consists of two parts (C1 and DenseBlock) used to extract deep features. C1 has 3 x 3 filters and DenseBlock (shown in Fig. 3) contains 3 convolution layers which also contains 3 x 3 filters. Number of channels in each convolution layer is 16. The filter size is 3x3 with stride of 1. The output of feature extractor is an input to the fusion layer. The output of the fusion layer is the input to the reconstruction layer. The image reconstruction layer has 4 convolution layers (3x3 filters). Dense block architecture has many advantages: 1) it preserves as much information as possible; 2) it can improve gradients and information flow; 3) it reduces overfitting. MFIF is a multi-class classification problem. $\mathrm{I_i}$ (*i*= *1, 2,..n*) is denoted as Input images captured at different focal offsets.

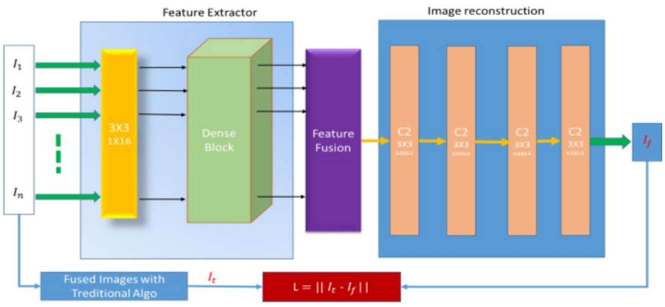


Figure. 2. Block Diagram of Network Architecture

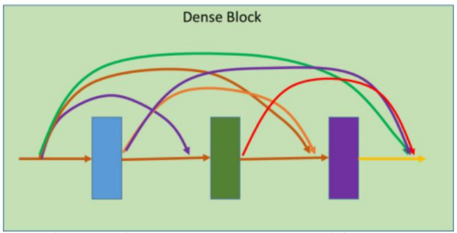
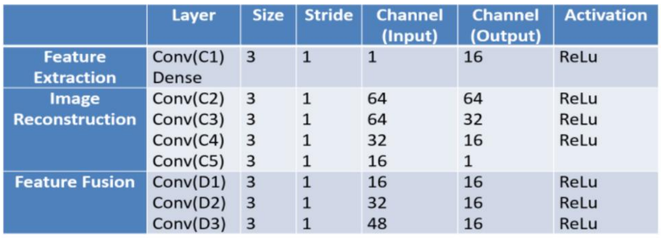


Figure. 3. Dense Block Architecture



	Layer	Size	Stride	Channel (Input)	Channel (Output)	Activation
Feature Extraction	Conv(C1) Dense	3	1	1	16	ReLu
Image Reconstruction	Conv(C2)	3	1	64	64	ReLu
	Conv(C3)	3	1	64	32	ReLu
	Conv(C4)	3	1	32	16	ReLu
	Conv(C5)	3	1	16	1	
Feature Fusion	Conv(D1)	3	1	16	16	ReLu
	Conv(D2)	3	1	32	16	ReLu
	Conv(D3)	3	1	48	16	ReLu

Table I. Architecture of the network

During the training process, our ground truth is the fused image (I_t) using traditional algorithms propriety of KLA-Tencor Corporation. Our main goal is to minimize the number of input images to reconstruct the fully focused image during inference. Thus, our loss function L is as follows:

$$L = ||I_f - I_t||_2 \quad (1)$$

Where I_f = Final fused Image (from the network)

And I_t = Ground truth Image (from KLA-Tencor's fusion algorithm)

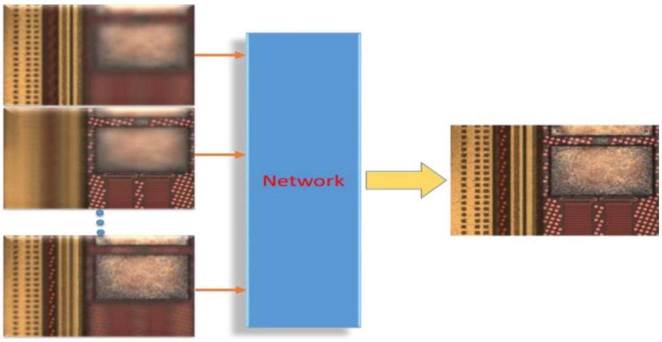


Figure. 4. Overview of the Network

B. Feature Fusion

Different fusion techniques can be used such as simple addition (Fig. 5.) or L1 Norm (Fig. 6.) etc. The output of the previous step is image D which is fused with the input images using weighted average rule to create the final fused map. Fused feature $f^m(x,y)$ can be calculated as follows:

$$f^m(x,y) = \phi_1^m + \phi_2^m \quad (2)$$

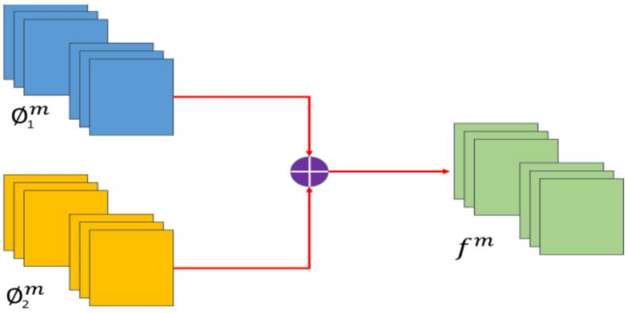


Figure. 5. Feature Fusion Strategy (addition)

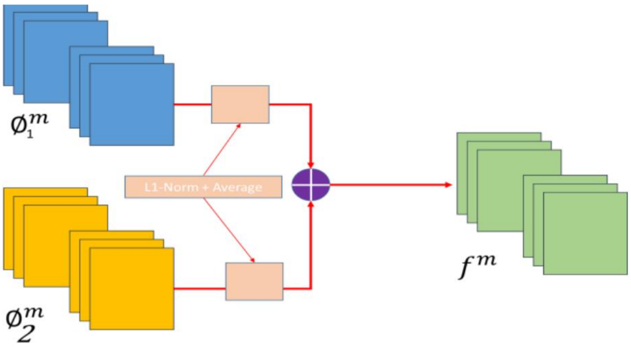


Figure. 6. Feature Fusion Strategy (L1 Norm)

Our model is designed to input 'n' number of multimodal defocused images and outputs the fully focused image by fusing focused features extracted from individual images. This helps in fusing images with wider range of focus offsets at a given site on a semiconductor wafer. In order to speed up transfer learning, the fusion layer is modularized such that at any given point of time in the future, a new feature fusion methodology can easily be integrated with our network to get a better result.

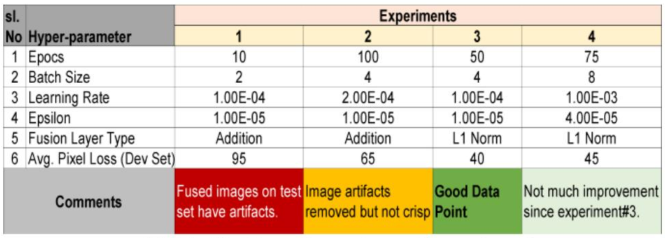
V. Experiments/Results/Discussion

A. Experiments & Network Standardization

The following experiments were performed on our network in order to standardize it.

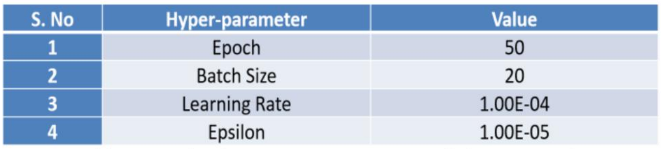
- We started with a publically available network and tested it with 2 semiconductor wafer input images and the results didn't look good.
- Experiment-1: We then added a fusion layer with a simple addition fusion and trained the network with hyper-parameters shown in Table II (Experiment 1) and the final image had artifacts.

- Experiment-2: We then trained using increased epoch size, batch size and learning rate (Experiment 2) but the resulting image wasn't crisp.
- Experiment-3: We then reduced the number of epochs, learning rate and changed the fusion layer from simple addition to L1 Norm (Experiment 3). Pixel loss reduced and the images looked good.
- Experiment-4: We then increased the number of epochs, learning rate and epsilon but the final image didn't show much improvement (Experiment 4).
- Finally, we settled with hyper-parameters shown in Experiment 3 as our network (shown in Table III).



sl.		Experiments			
No	Hyper-parameter	1	2	3	4
1	Epocs	10	100	50	75
2	Batch Size	2	4	4	8
3	Learning Rate	1.00E-04	2.00E-04	1.00E-04	1.00E-03
4	Epsilon	1.00E-05	1.00E-05	1.00E-05	4.00E-05
5	Fusion Layer Type	Addition	Addition	L1 Norm	L1 Norm
6	Avg. Pixel Loss (Dev Set)	95	65	40	45
Comments		Fused images on test set have artifacts.	Image artifacts removed but not crisp	Good Data Point	Not much improvement since experiment#3.

Table II. Experiments on the network



S. No	Hyper-parameter	Value
1	Epoch	50
2	Batch Size	20
3	Learning Rate	1.00E-04
4	Epsilon	1.00E-05

Table III. Final Hyper-parameters of the network

After about 50 epochs, our network converged to a decent average pixel loss of ~25 gray levels (training set) and ~40 gray levels (dev set). Even after training further, the convergence didn't improve. Performance on our dev set data seems to be diverging after about 65 epochs (See Fig. 7). We standardized our network at 50 epochs. Our test set was evaluated on the standardized network with roughly 50 sites out of which about 2 sites did not perform well. One example of a good and bad final fused image is shown in Fig. 8.

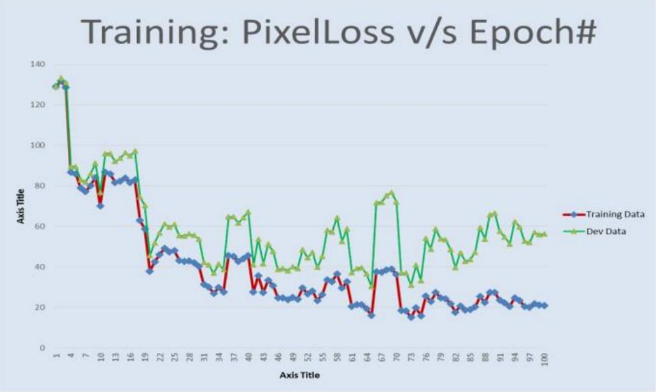


Figure. 7. Performance of Train/Dev set

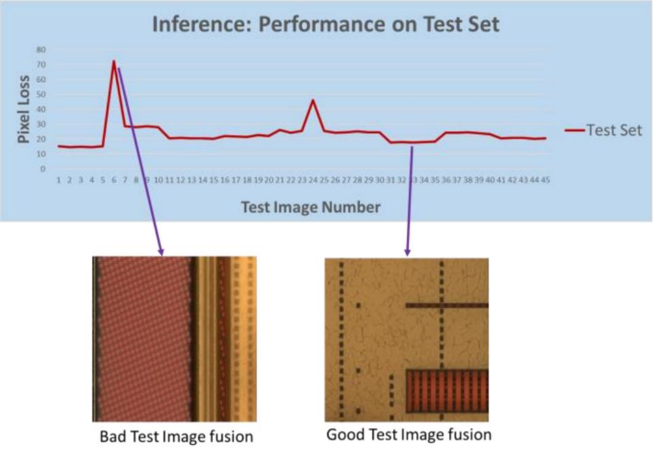
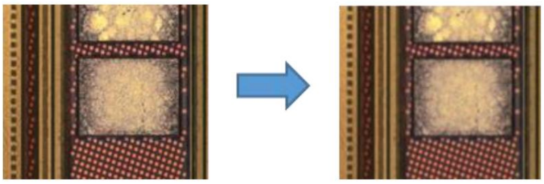


Figure. 8. Performance of Test set

B. Fusion results analysis

We analyzed the performance of the network by computing the R, G, B channel image difference as shown below (Fig. 10.). The gray level difference between the Ground Truth image and output image is about 25 gray levels which is what we see on an average in Fig. 7. as well.



Ground Truth Image Output Image

Figure. 9. Ground Truth and Output Image

Figure. 10. R/G/B Difference Image (GroundTruth – Output)

C. Patch-Based Image fusion

Our test data on some data sets showed a high gray level value (about ~85 gray levels). This led us to explore different methodologies to apply different image fusion techniques one of which was the patch based image fusion.

The input image is decomposed into multiple patches (configurable patch size 4X4, 16X6, 32X32 etc.) and fed the batch sets to the network to construct the focused patch for each location. These patches are stitched together to construct final focused image as shown in the Fig. 11. below.

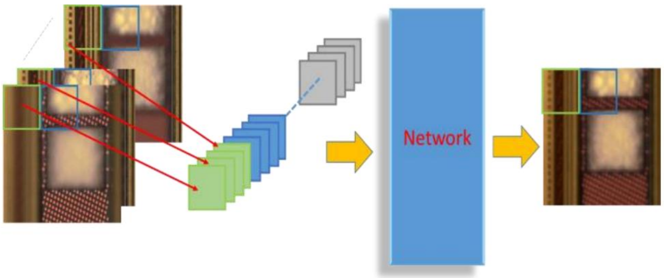


Figure 11: Patch based image fusion

The patch-based image fusion technique worked well with individual patch level (visually the features inside the Patch look good), but the overall image look noisy as the transition across the patch boundaries were not smooth (see Fig. 12). To get rid of patch transition artifacts, we need to implement overlapped patch transition which is part of the future work and beyond the scope of this paper.

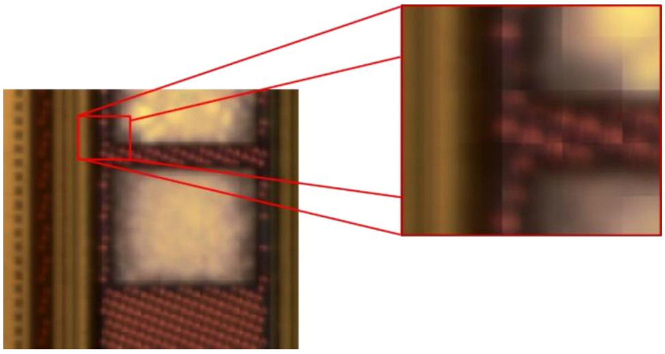


Figure 12: Patch based image fusion Noisy path boundary

One observation we made is the fusion performance improved as the patch size decreased but the network took longer to generate the final output. Results on different patch size are shown in Fig. 13.

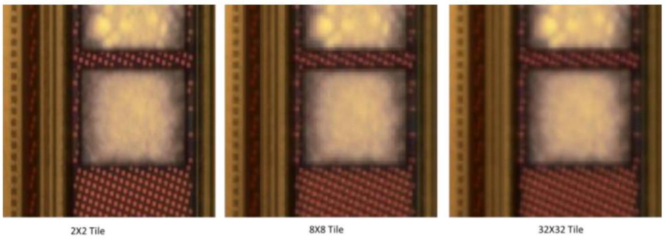


Figure 13: Fusion performance with different patch size

VI. Conclusion/Future Work

We were able to extend an existing architecture to use multiple input images taken at different focal length to generate a fully focused image. Performance of our network on a site with high density features, degrades slightly.

Currently our feature fusion strategy is a simple addition and L1 Norm. Future work includes a better feature fusion methodology to improve the performance on different semiconductor wafer images including patch based image fusion using overlapped patch transition.

VII. Contributions

Srinivasa Rao worked on the different Image fusion techniques. He also worked on converting the network to patch based architecture.

Antariksh De worked on changing the network to take multiple images instead of two. The experiments for various hyper-parameters were performed by Antariksh.

We take this opportunity to thank Ankur Gupta who works for KLA-Tencor. He was instrumental in helping us collect the input and ground truth data. His expertise in image fusion immensely helped us through our project.

Antariksh De, Srinivasa Rao and Ankur Gupta together worked on the research papers and collecting vital information prior to starting the project.

REFERENCES

[1] Y. Liu, X. Chen, H. Peng, and Z. Wang, "***Multi-focus image fusion with a deep convolutional neural network***," Information Fusion, vol. 36, pp. 191–207, 2017

[2] H. Tang, B. Xiao, W. Li, and G. Wang, "***Pixel convolutional neural network for multi-focus image fusion,"*** Information Sciences, Vol 433- 434, pp 125 – 141, 2017.

[3] X.Yan, SZ. Gilani, H. Qin, A. Mian, "***Unsupervised Deep Multi-focus Image Fusion",*** arXiv preprint arXiv:1806.07272, 2018

[4] C. Du, S. Gao, "***Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network",*** IEEE Access, 5 (2017), pp. 15750-1576

[5] D. Guo, J.Yan, X.Qu, "***High Quality multi-focus image fusion using self-similarity and depth information***", Opt. Commun. 338(1) (2015) 138- 144

[6] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "***Pixel-level image fusion: A survey of the state of the art***," Information Fusion, vol. 33, pp. 100–112, 2017

[7] S. Li, X. Kang, and J. Hu, ``***Image fusion with guided filtering,***'' IEEE Trans. Image Process., vol. 22, no. 7, pp. 2864_2875, Jul. 2013.

[8] Liu, S. Liu, and Z.Wang, ``***Multi-focus image fusion with dense SIFT***,''Inf. Fusion, vol. 23, pp. 139_155, May 2015B.

[9] Yang, S. Li, "***Multifocus image fusion and restoration with sparse representation***", IEEE, Trans. Instrum. Meas. 59 (4) (2010) 884–892