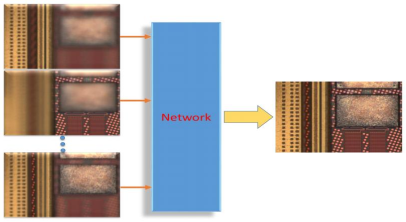
Multi-focus image fusion with a Deep Convolutional Neural Network for Semiconductor Wafer Inspection tools

Srinivasa Rao (srao78@stanford.edu) and Antariksh De (antde@stanford.edu)

Motivation

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.



Data

The data set was collected using multiple customer wafers at multiple focal offsets and at various high topology feature locations, 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.
- 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.
- 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.

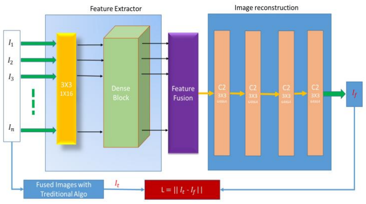
Features

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.

Network Architecture

Our proposed network is Model 1, however we also evaluated Model 2 shown below.

Model 1



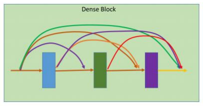
The goal was to minimize the number of input images to reconstruct the fully focused image during inference. Thus, our loss function L is as follows:

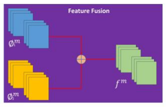
$$L=||I_f - I_t||_2$$

Where I_f = Final fused Image (from the network)

And I_t = Ground truth Image

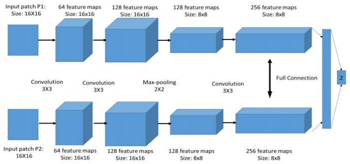
S. No	Hyper-parameter	Value
1	Epoch	50
2	Batch Size	20
3	Learning Rate	1.00E-04
4	Epsilon	1.00E-05





	Layer	Size	Stride	Channel (Input)	Channel (Output)	Activation
Feature Extraction	Conv(C1)	3	1	1	16	ReLu
	Dense					
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

Model 2

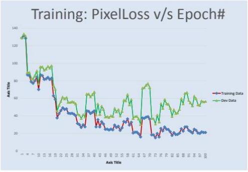


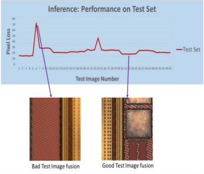
S. No	Hyper-parameter	Value
1	Epoch	10
2	Batch Size	128
3	Momentum	0.9
4	Weight Decay	5.00E-4
5	Learning Rate	1.00E-04

Results

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 (as shown in the left figure below). 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 the right figure below:

Youtube Link for Poster Presentation: https://www.youtube.com/watch?v=I_oUuROuM94&feature=youtu.be





sl. No	Hyper-parameter	Experiments 1	Experiments 2	Experiments 3	Experiments 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.

Discussion

Our model (Model 1) 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.

Model 2 is designed for Bi-modal image fusion(background and foreground) and can't be extended for multi-modal image fusion. This model works on small image patches which has an adverse affect on the memory consumption and computational efficiency. These shortcomings have been addressed in our network (Model 1).

Future

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.

References

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

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

[3] 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

[4] B. Yang, S. Li, "**Multifocus image fusion and restoration with sparse representation**", IEEE, Trans.