Improving Automated Visual Fault Detection by Combining a Biologically Plausible Model of Visual Attention with Deep Learning

Frederik Beuth [ID], Tobias Schlosser [ID], Michael Friedrich [ID], and Danny Kowerko [ID]

Junior Professorship of Media Computing,
Chemnitz University of Technology,
09107 Chemnitz, Germany,

frederik.beuth@cs.tu-chemnitz.de   tobias.schlosser@cs.tu-chemnitz.de
michael.friedrich@cs.tu-chemnitz.de   danny.kowerko@cs.tu-chemnitz.de

Abstract—It is a long-term goal to transfer biological processing principles as well as the power of human recognition into machine vision and engineering systems. One of such principles is visual attention, a smart human concept which focuses processing on a part of a scene. In this contribution, we utilize attention to improve the automatic detection of defect patterns for wafers within the domain of semiconductor manufacturing. Previous works in the domain have often utilized classical machine learning approaches such as KNNs, SVMs, or MLPs, while a few have already used modern approaches like deep neural networks (DNNs). However, one problem in the domain is that the faults are often very small and have to be detected within a larger size of the chip or even the wafer. Therefore, small structures in the size of pixels have to be detected in a vast amount of image data. One interesting principle of the human brain for solving this problem is visual attention. Hence, we employ here a biologically plausible model of visual attention for automatic visual inspection. On this basis, we propose a hybrid system of visual attention and a deep neural network. As demonstrated, our system achieves among other decisive advantages an improvement in accuracy from 81 % to 92 %, and an increase in accuracy for detecting faults from 67 % to 88 %. Therefore, the error rates are reduced from 19 % to 8 %, and notably from 33 % to 12 % for detecting a fault in a chip. Hence, these results show that attention can greatly improve the performance of visual inspection systems. Furthermore, we conduct a broad evaluation, which identifies specific advantages of the biological attention model in this application, and benchmarks standard deep learning approaches as an alternative with and without attention.

This work is an extended arXiv version of the original conference article published in "IECON 2020". It has been extended regarding visual attention, covering (i) the improvement and equations of visual attention model, (ii) a deeper evaluation of the model, (iii) a discussion about possibilities to combine the attention model with the DNN, and (iv) a detailed overview about the data.

Index Terms—Semiconductor Manufacturing, Factory Automation, Fault Inspection, Wafer Dicing, Laser Cutting, Computer Vision, Deep Learning, Convolutional Neural Networks, Visual Attention

I. INTRODUCTION

One long-term goal is to incorporate biological processing principles into machine vision systems. Visual attention, a smart human processing principle that focuses processing resources on an aspect of a scene relevant for the current task, is one of these biological processing principles [1, 2]. We apply this principle here to the domain of wafer dicing to investigate its benefits and under which circumstances it improves automated visual inspection systems. The long-term goal of our research is therefore to better understand the power of human processing as well as to incorporate its benefits into machine vision systems and improve them accordingly.

A major aim in the domain of the semiconductor industry is to detect and recognize production errors and faults early on. As manual detection is a very labor-intensive and thus costly procedure, computer vision systems are often deployed as an automatic detection system [3, 4]. This does not only results in reduced manufacturing costs and work load, but also helps increasing the yield of the production process itself. Hence, systems for automated visual inspection are widely deployed in the industry.

In this contribution, we address the topic of wafer dicing. Wafer dicing is the separation of silicon wafers into single components, e.g. chips, often using a dicing saw [5, 6]. Dicing based on laser technology is a novel alternative method to separate brittle semiconductor materials via thermally induced mechanical forces (laser cut wafer dicing [5, 6]). Thereby, a dicing street is the area where dicing is potentially allowed. The quality criterion of dicing is that the laser cut (kerf) must not leave the street (Fig. 1). A curve leaving a street entails faulty chips and decreases the wafer yield, i.e. the ratio of faultless to the total sum of chips.

The field of automated visual inspection for wafer dicing uses classically image processing approaches commonly differentiated due to their functionality in projection-, filter-based, and hybrid approaches [3, 7]. Projection-based approaches include for example principal component analysis, whereas filter-based approaches encompass spectral estimation methods, yet, they often need manual adaption. Therefore, learning-based and hybrid approaches make utilization of support vector machines (SVMs) or multilayer perceptrons (MLPs) [8, 9], while, in recent years, also a few more powerful deep learning (DL, [10]) approaches have been deployed [11]–

Figure 1: Overview of a wafer (left) with chips (middle) as well as faulty and fine streets (right) as a result of the wafer dicing process. (The shown examples were generated synthetically from the original images to protect the intellectual property, while retaining a close resemblance to the original imagery).

[15], which we will address here. For instance, [14] tested synthetic wafer data using a convolutional neuronal network (CNN) in their application. Others recognize the production process over time for an on-the-fly control [12], or use recurrent neuronal networks [13].

However, one problem is that faults are typically very small in size and hard to detect in the large wafer disk (Fig. 1). Imaging systems capture the complete (often stitched) image of a wafer, which results in resolutions of up to 150 megapixels, where single chips often range in a size of 200 – 2000 pixels as in our case. Faults are even smaller structures up to the size of only a few pixels (Fig. 1, 2). Therefore, the challenge is to detect these small structures in a vast amount of data. Classical deep neural networks have trouble to deal with this problem: Either the input image is down-scaled as normally for the network, but then the small structures would be lost. Or, the input could be chosen large enough, but then the network would run very slowly and had too many free parameters, leading to overfitting problems, as the network's size would increase unfeasible.

One interesting biological processing principle for this problem is visual attention, a smart human mechanism to select from the huge amount of input data the relevant one for the task at hand [16], or to focus neuronal processing resources on an aspect of a scene [1, 2]. The principle is deployed here to "zoom in" into the wafer disk or into a chip (Fig. 1). Visual attention has, to the author's knowledge, not been used in the domain yet. Therefore, we like to propose the first model of visual attention, in combination with deep learning, for the domain of fault recognition in the semiconductor industry.

There exist several approaches of combining visual attention with deep neural networks (DNNs) in other domains already. The background behind this idea is often that attention focuses the processing on a part or an aspect or a scene, which is then passed for classification to a machine learning classifier, like a DNN. In such a way, the attention model zooms in on the content, which then the DNN classifies in higher resolution and with less irrelevant data. However, the way to combine this remains pretty unclear and under debate. We found several dominant approaches, which are listed in the following. 1) One

Figure 2: Our data material is heterogeneous, and it consists of several different wafers, for which exemplary streets are depicted.

approach is the saliency models [17]. The idea understands attention as a mechanism to define a spatial region (region of interest, ROI [17]), which is then later passed to a classifier. Newer work uses this idea for taking fast snapshots from the image via attention and feed them to a CNN (Glimpse approach [18]). 2) As saliency models struggle with objects that are not very obvious in the image, they find the wrong regions for the DNN. This has led to a combination of this approach with machine learning like reinforcement learning [19], e.g. [20]. 3) Similarly, a saliency model can be controlled by words to find the visual regions corresponding to text (attention for visual-textual alignment, e.g. Wang et al. [21]). 4) A different approach is from Jürgen Schmidthuber [22], which introduces top-down feature-based attention in his model and thus uses attention towards specific features, not regions. 5) Attention towards some basic features can also be used to select an irregularly shaped region, defined via visual features instead of a spatial region, and then feed this content to a DNN (e.g. [23]). 6) In the recent few years, works also utilize DNNs, or building blocks of them, as attention networks. So, one neuronal network serves as attention network, while the other is modulated by it and processes normally the image (e.g. [23]–[25]). Quite a few systems use this approach, but they are designed very diverse (5 examples: [23]–[27]).

Therefore, many different approaches exist that allow the combination of visual attention with machine learning and deep learning based models, yet, selecting the approach with the best performance for a specific task remains difficult.

However, when we look at the attentional processing in the brain (e.g. Miller & Buschman [28] or Tsotsos et al. [29]) and the underlying connectivity (e.g. Ungerleider et al. [30] or Felleman & Van Essen [31]), we found that none of the above approaches resembles closely the attentional processing in the human brain. Also none of these approaches are underpinned with much neuroscientific data. For instance, there exist a lot of single-cell recordings in the attention literature [32, 33] which are not replicated at all, and also many behavioral influences by attention [16]. The saliency models are inspired by the latter [17], but the approaches are nowadays also not so close more linked to psychology. Moreover, there exist also many other different ways in which visual attention alters human behavior, which are not reflected (e.g. texture segmentation [34, 35]).

Therefore, we propose here to use a more biologically plausible model of visual attention, to avoid all of these problems, and the quarrels about which model is the right one. Biologically-plausible models of visual attention originate more from the discipline of computational neuroscience, which develops models of the human brain to replicate neuronal recording data and to simulate human behavior. In recent years, a few works have also shown that such models of visual attention are capable of real-world applications (see dissertation [2] for an overview), e.g. [1, 2, 29, 36]–[38]. However, none of these models, at least to the authors' literature search, have been combined with deep learning yet.

We will employ the model of Beuth, 2019 [2] as it shows very promising real-world capability (see also [39]) while maintaining a deep biological plausibility. The model has a strong neurophysiological foundation as it can replicate a large range of neuronal recordings of attention [33]. Moreover, it has a satisfying operation as several applications show it can at least deal with virtual reality [40] and real-world [39]. And finally, it can replicate human behavior as it is based on previous models (visual search, [1, 41]) and is able to also fit new experiments in other behavioral paradigms (OSM, Chap. 5 in [2]). Additionally, the work [2] shows advantages over saliency models: a) Top-down object-descriptors, b) biological foundation, c) the option to easily realize complex task sets in a natural way, as later shown in Sec. II-B. The current work is deeply rooted in neuroscience and focuses on the concept of visual attention, while extending an earlier work-in-progress publication of us using opposingly a specialized pipeline [7]. The current biologically-grounded contribution also evaluates, in comparison to this shorter work, the concept of visual attention with deep learning more broader and in-depth. This is more thoroughly possible as we utilize now a model based on the human brain. Therefore, we like to propose the combination of a biologically-plausible model of visual attention with deep learning as our second contribution.

II. PROPOSED SYSTEM

We design a multi-stage system (Fig. 3), which processes a whole wafer to detect faulty and faultless chips. Wafer images are recorded via different microscopes, either in the

Figure 3: Overview of the whole system.

form of unstitched subimages or, alternatively, preprocessed via the microscope software by stitching all subimages into one image. In the system, the chips are at first classified into chips inside the wafer area and chips on the wafer border. The latter are also scanned by the microscope, but are typically incomplete or broken (Fig. 1, wafer border). The manufactures are not much interested in faults in the outer chips, thus we only process the inner chips further.

Afterwards, the visual attention model is employed to find the region of interest (ROI) for the inside chips. Regions of interest are in our application the chip borders, streets, and their surroundings. The ROIs are then passed to the CNN for detecting faults in the regions. Finally, we calculate if a chip is faulty or not, depending on the classification of the four borders of each chip. To show the benefits of attention, we will compare our full system to approaches without attention and to approaches of the automated visual fault detection.

A. Classification of inside vs. wafer-border chips

This first stage distinguishes chips inside the wafer from chips on the wafer border. The border chips are typically broken so they have to be labeled for the user as "wafer border chip" and are excluded from evaluation. However, the recognition is pretty easy as the pattern is very obvious. The chip is crossed by a large black shape constituting the wafer border (Fig. 1). We found, that a basic CNN is already able to achieve an accuracy of over $99\,\%$, utilizing two convolution layers, one max pooling, one dense and one fully connected layer. This corresponds to a VGG network [42] with only one block.

B. Model of visual attention

Our attention model (Fig. 4) stems from a line of computational-neuroscience models of visual attention from the group of Fred Hamker [1, 2, 33, 39, 40, 43]. The model is adapted and further developed from [2, Chap. 4].

In the attention model, at first, the input image is processed through an earlier visual brain area (V1), which filters the image for edges and colors (step (i) in Fig. 4). Edges are recognized by Gabor filters, and colors as a color contrast between red and green, and between blue and yellow [2]. Colors are not used much in this wafer application, but are included for the generality of the model. This stage models brain areas like the lateral geniculate nucleus (LGN) and the primary visual cortex (V1).

Figure 4: The visual attention model in the task to search for a street. The input image shows a chip of wafer 3, whereby again the image was abstracted. In the higher visual area (HVA), each box shows the activity of a street neuron in image coordinates, whereby what pattern the neuron detects is visible in the inlet. The red circle marks the center position of the found target. See the main text for the area abbreviations. The chip's borders are marked with a dotted green line.

The neuronal responses are then routed to a higher-level visual area (HVA), encoding objects (ii). This area simulates high-level visual areas in the brain such as the inferior temporal cortex, where these cells have been found [44]. The responses of the model's HVA neurons are computed by convolving the neurons of V1 with a pre-learned weight matrix. The weight matrix can be learned by any offline learning method, we use a simple one which was already employed in previous research, the one-shot learning [2, 40]. The procedure learns an object directly from a single image, leading to its name. We apply it here to learn cells reacting to the chip borders. We learned 12 different cells to account for the variance in the data material: We combine 2 street orientations (vertical and horizontal), 3 street widths, and 2 color schemes (black streets/white chips and vice versa). Four cells are exemplarily shown in Fig. 4, no. 1, 3, 4 and 6.

The model has to search in this task for the borders of a chip (Fig. 4). This task is called visual search in psychology and it is known that top-down visual attention is applied to the search target [45], realized in the model by signals from the prefrontal cortex (PFC), encoding this task instruction, to the higher

visual area (iii). These signals amplify neurons encoding the edges in HVA, leading to higher activity at potential edges in the image. This realizes the human task instruction "check the streets of the chip", which we know from human workers is carried out by looking at the chip border regions (streets).

Afterwards, the HVA activities are spatially processed by a brain area called frontal eye field (FEF [2], iv-vii). The FEF first takes the maximum over the HVA activity, the result shows activity on potential target object locations. Afterwards, a competition between places is applied and the activity is projected back to HVA, forming a reentrant loop (vii). This loop focuses neuronal activity to a single location over time (spatial attention). When a single location is selected in FEF movement (FEFm), it denotes an upcoming eye movement. In humans, the eye movement would select the target location, thus we simply read out it and define it as street center (viii).

Which of the four streets is selected is more or less random. After the first street is selected, the location is suppressed in the FEF by an inhibition-of-return standard approach (IOR, [1]) via the external attention signal, so another street will be selected. This task is repeated 4 times. Finally, 4 street regions are cut out based on the found center coordinates of each street.

The model was further extended from previous work for this application by: introducing the IOR concept from previous attention models [1], adding an external attention signal, and increasing the precision of the street center. The changes are listed in the Sec. II-D. In general, the model is described by a set of differential equations [2].

C. Advantages of the attention model

Besides the general advantage of visual attention, we have analyzed our work regarding the advantages of such a biological attention model in the current application.

First, the model searches a street pattern based on high-level object descriptors and not merely pixel-based. The object is encoded by a neuron in the higher visual area (Fig. 4). Multiple different streets can be encoded by multiple object templates, thus also a certain variance can be encoded. The model then searches for all templates in parallel. The high-level object descriptors make the model more robust compared to classical computer vision approaches like edge following, e.g. against noisy areas or stitching errors, while keeping the model simpler than deep learning solutions.

Due to the nature of the employed learning, i.e. the one-shot learning, the method learns a template directly from a single image. Hence, it is very fast in its nature and the total runtime of the learning is only a few seconds on consumer-grade hardware. The other advantage of the learning procedure is that the image can be defined in a rather conceptual manner, more of a sketch than an actual image (Fig. 4). We found that even defining the image via simple image editor software is enough. Hence, it can be easily and swiftly produced.

As a last advantage, the attention model can deal with the inner structures of the chips out of the box by an external spatial attention signal. This signal defines which chip areas

are suppressed regarding processing. We use it to suppress the inner structure of a chip, as we know human inspectors would also not look at the middle of a chip image. To illustrate this, we use a chip with especially a lot of inner structures in Fig. 4. The suppression map can again simply be defined as an image. The model naturally realizes in this way task instructions. During the processing, the initially very noisy activity is filtered, illustrated by showing the input stages towards the FEF (iv,v,vi): the first stage is very noisy as it reacts to all inner structures of the chip (iv). Yet, the external spatial attention signal is then applied to this, which suppresses the "middle" by decreasing neuronal activity (v). Afterwards, the competition takes place, increases the signal contrast, and reduces the activity to a few locations (vi). During this processing, the model filters out inner chip structures, which might be very similar to the searched streets and thus would divert the recognition process.

D. Improvements of the visual attention model

The attention model has been modified in several ways for this work. Additionally, as the attention model stems from a line of older attention models, this section provides the changes as compared to the previous works [2].

1) Spatial external attention signal: The model was extended with an external spatial attention signal. In the data set, some of the chips have line-like structures within the chip. We know from human workers that they were told to ignore the structures inside a chip, which translates roughly to the task instruction 'do not look in the middle of a chip'. Such task instructions can be naturally realized in the model by directing 'negative' attention to the middle of a chip.

Hence, this signal realizes spatial components of task instructions. We implemented this external attention signal as an incoming signal to the frontal eye field (FEF), as this structure is highly involved in spatial attention [41, 46].

The equations in the FEF were changed in the FEF visual cells from the ones in [2] to the following ones by adding the term of $[\cdot (1 + 2 \, a^{\text{FEF}})]$:

$$\tau^{\text{FEFv}} \frac{\partial r_x^{\text{FEFv}}}{\partial t} = -r_x^{\text{FEFv}} + E_x \tag{1}$$

$$\text{with:} \quad E_x = C \left(Q \left(F_x \right) \right) \tag{2}$$

$$F_x = [E_x^{\text{HVA2}} \cdot (1 + 2 \, a_x^{\text{FEF}})]_0^1 \tag{3}$$

whereby a^{FEF} denotes the spatial attention signal, resulting in a neuronal activity matrix of the same size as the FEF. The entries are $a_x^{\text{FEF}} \in [-0.25, 0.25]$ and thus can be negative to allow a 'negative' attention signal. The variable E_x^{HVA2} denotes the incoming signal from the HVA. In the doctoral thesis [2], the FEF received also an incoming signal from V1 (E_x^{V1}), which is disabled here as it is not used and thus set to 0 (equations number Eq. 4.53 - 4.55 in [2]). The helper functions $C(x)$ and $Q(x)$ denote a non-linearity (C) to increase the difference between low and high signals (Eq. 4.59 in [2], not listed here), and a signal enhancement operation (Q, Eq. 4.60 in [2]).

2) Inhibition of return (IOR): To detect multiple streets in the chip image, an inhibition-of-return approach (IOR [1, 17]) was also added to the model. IOR describes the concept that the target region is inhibited after an eye movement, so the next eye movements do not visit this region again. This allows the search and localization of several streets in a row within a chip image. The IOR is implemented in our model by an IOR map, which stores locations of previous eye movements as activity blobs and suppresses then the FEF cells (standard approach, [1, 17]).

The idea was realized as follows: We define an inhibition map (r^{IOR}), initialized initially with a small positive number uniformly (0.25), representing no inhibition. The image is then processed through the normal duration of the model. When an eye movement (saccade) is executed, inhibition at this position x_s is set as a blob of negative values (Eq. 4, 5):

$$G = \mathfrak{g}(x_s, 1, \frac{\#FEF_{x,y}}{6}) \tag{4}$$

$$r_x^{\text{IOR}} = r_x^{\text{IOR}} - v^{\text{IOR}} \cdot G, \tag{5}$$

whereby:

- The parameter $v^{\text{IOR}} = 0.75$ denotes a scaling of the IOR influence.
- The function $\mathfrak{g}$ represents a two-dimensional Gaussian function centered at $(0,0)$, whereby a denotes the amplitude, and σ the standard deviation:

$$\mathfrak{g}(x, a, \sigma) = a \cdot \exp \left(- \left(\frac{(x_1 - 0)^2}{2\sigma_1^2} + \frac{(x_2 - 0)^2}{2\sigma_2^2} \right) \right)$$

Afterwards, the image is processed again, and the map r^{IOR} inhibits the FEFv cells on the previous locations. This inhibition leads to the selection of a different target, i.e. a different street. The IOR influence is also realized as a modulatory influence into the FEF, like a 'negative' attention signal, and is bundled with the other attention signal, the external one, via a Fuzzy-Min operation [47]. With the IOR component, the Eq. 3 changes to the following Eq. 6:

$$F_x = [E_x^{\text{HVA2}} \cdot (1 + 2 \cdot \min \{a_x^{\text{FEF}}, r_x^{\text{IOR}}\})]_0^1 \tag{6}$$

3) Precision of the attention model's localization: First of all, we increased the spatial precision of the attention model by three improvements:

A) Adapting V1. V1 consists of two layers, V1 simple and pool, whereby the V1 pooling layer has a lower spatial resolution $(1 : 10)$. We improved the behavior of the model by aligning the two layers on top of each other, implying that V1 pool must have precisely $^1/_{10}$ of the size of V1 simple.

B) The behavior of the soft-max pooling in HVA layer 2/3 was changed in such a way that it considers high activities in the input (HVA layer 4) more strongly. HVA consists of two layers, HVA layer 4 and HVA layer 2/3, simulating the layers of a single cortical brain area [33], whereby layer 2/3 is a pooling layer. For this change, the non-linearity in the soft-max operation was adapted from $p1 = 4, p2 = 1/4, v^{\text{HVA4}} = 1$ to $p1 = 8, p2 = 1/4, v^{\text{HVA4}} = 16$ (Eq. 7, corresp. to Eq. 4.50 in [2]).

Unit	Layer	Type	Output shape	Kernel size	Stride
conv1	conv1_1	conv	$56 \times 188 \times 32$	5×5	1
	conv1_2	conv	$54 \times 186 \times 48$	3×3	1
	pool1	max pool	$18 \times 62 \times 48$	3×3	3
	dropout1	dropout	$18 \times 62 \times 48$	/	/
conv2	conv2_1	conv	$16 \times 60 \times 64$	3×3	1
	conv2_2	conv	$14 \times 58 \times 96$	3×3	1
	pool2	max pool	$7 \times 29 \times 96$	2×2	2
	dropout2	dropout	$7 \times 29 \times 96$	/	/
conv3	conv3_1	conv	$5 \times 27 \times 144$	3×3	1
	conv3_2	conv	$3 \times 25 \times 192$	3×3	1
	pool3	max pool	$3 \times 8 \times 192$	1×3	1×3
	dropout3	dropout	$3 \times 8 \times 192$	/	/
fully conn	dense1	fully conn	192	/	/
	dropout4	dropout	192	/	/
	dense2	fully conn	2	/	/

Table I: The CNN for street classification.

$$E_{d,i,x} \;=\; \left(v^{\text{HVA4}} \cdot \sum_{x' \in \text{RF}} w_{x'}^{\text{HVA4-HVA2}} \left(r_{d,i,x'}^{\text{HVA4}} \right)^{p_1} \right)^{p_2} \quad (7)$$

$$w_{x'}^{\text{HVA4-HVA2}} \;=\; \mathfrak{g}(x', 1, [1,1]) \quad (8)$$

C) The size of HVA layer 2/3 and the FEF was increased to have a finer resolution for selecting the target. Normally, HVA layer 2/3 has a lower resolution than HVA layer 4 to decrease the spatial information over the visual layer hierarchy. However, as also the FEF has the same resolution as HVA layer 2/3 (assumption in the original model), we increased the resolution in this area to have also a higher resolution in the FEF, hence resulting in more accurate eye movements.

4) Equations adapted to the task: Furthermore, several additional parameters were adapted. They were changed as the attention model was applied to a new application input, which causes different strong and broad activities in the neuronal layers.

- The signal HVA layer 4→HVA layer 2/3 was not only increased due to the changed non-linearity (see previous paragraph, 3-B), but also to account for a general lower HVA layer 4 activity: $v^{\text{HVA4}} = 1 \rightarrow 16$.
- The signal FEFvm→HVA layer 4 was decreased to account for a broader FEF activity: $v^{\text{FEFvm-HVA4}} = 4 \rightarrow 3$ and $v^{\text{SP-1}} = 0.85 \rightarrow 0.3$ (Eq. 4.37, 4.44 in [2]).

E. CNN for street classification

The streets were then classified by a convolutional network, as outlined in Tab. I, in 3 classes. 'Good streets' (Class 0) represents chip regions without a fault, and 'bad streets' (2) with a fault. Additionally, we defined a class 'anomalies' (1), describing intact streets, but with an unknown visual event on them. The plan is to report back these chips in the later production system to a human worker for a further manual inspection. For the analyses in this paper, this class is not considered, and its samples are put into the 'good streets'.

The attention model returns the center where it has found a street. We define a region of interest (ROI, [17]) of $120\,\%$ the size of the chip and $6\times$ the width of the street around this center. The goal of our application is to detect dicing faults, which occur in the space between the streets and the chips,

and possibly continue inside the chip. To cover these areas at best, the ROI was centered in such a way that the main street is placed at $1/3$ of the image height such that $2/3$ of the image shows the associated chip. Hence, the input street image contains the street itself as well as parts of adjacent chips and street crossings (Fig. 1, right side).

Regarding the design of the CNN, we indicate for the CNN the relevant chip areas inside the provided image and thus what part of the image should contain no cuts, by employing a trick from deep reinforcement learning [48]. The street regions are rotated before being passed on to the CNN, whereas the chip area is always located in the same image position (chosen as top). Thus, the position of the chip region is stable over all images in the data set. The trick is now to reduce spatial pooling to allow the CNN to learn specific weights for specific regions in the provided street image. As the distinction lays predominantly in the y direction (chip is on top, street on bottom), we lower the pooling in the direction of y. Following this principle, the effect of the pooling in y is reduced within the deeper pooling layers as shown in Table I.

F. Chip classification

Finally, the results of the street classification are translated back to determine faulty and good chips. If a chip has at least one faulty side, it is defined as faulty, otherwise, it is defined as good. The chip classification is the final outcome of the system.

III. TEST RESULTS AND EVALUATION

In this evaluation, we analyze the performance of our whole system, and quantify the improvements which result from employing visual attention. Yet, beforehand, we will validate if and how well the attention model operates in this task.

A. Data overview and test configuration

The wafer data set originates from a real-world, laser-based dicing process of multiple types of semiconductor wafers. During the process, each wafer was first mounted on a tape and a frame. Subsequently, the dicing tape was expanded in order to broaden the cut, making it possible to visualize the dicing streets using a wide-field light microscope. The used microscopes have a scanning stage which scans the wafer line-wise over n lines and m columns, allowing imaging of up to 150 mm / 6 inch wafers. The recorded subimages are stored separately or are stitched together, and then the images are saved offline.

The data set consists of 10 different wafers as more wafer material was not available due to copyright restrictions. A more detailed breakdown of our data set is shown in Tab. II. The wafers belong to 6 different types and thus are pretty heterogeneous (Fig. 1, 2). It is visible from the data that the samples per wafer are seriously imbalanced as the wafers exhibit different sizes and thus different numbers of chips in reality. This all must be considered by a real-world machine learning program.

Wafer	1	2	3	4	5	6	7	8	9	10	Total
Wafer type	1	2	3	3	3	3	1	4	5	6	
No. chips	744	5 007	566	566	566	566	664	144	728	2 050	11 601
No. inside chips	550	3 344	432	434	430	432	536	112	586	1 336	8 192
Class 0 – good	2 017	12 150	1 682	1 539	1 556	1 638	1 882	327	1 789	5 003	29 583
Class 1 – anomaly	90	239	27	143	144	64	151	78	421	12	1 369
Class 2 – bad	90	554	15	51	18	23	110	39	109	313	1 322

Table II: Number of samples in the data set, split into the numbers for each wafer, and for chips inside the wafer area. Additionally, the number of streets are given per wafer and per class.

Wafer	1	2	3	4	5	6	7	8	9	10	Total
Total streets	2 200	13 376	1 728	1 736	1 720	1 728	2 144	448	2 344	5 344	32 768
Found streets	2 197	12 943	1 724	1 733	1 718	1 725	2 143	444	2 319	5 328	32 274
Accuracy in %	99.86	96.76	99.77	99.83	99.88	99.83	99.95	99.11	98.93	99.70	98.49

Table III: Accuracy of the visual attention model to find the streets, given per each wafer.

Training of the CNNs was performed with the deep learning framework Keras and TensorFlow in Python. We split the data into a 50 % training, 25 % validation, and 25 % test set. All data were resized to 192×60 in advanced for the streets (and to 96×96 for the chips). We choose to have a rather small input size to keep the number of weights, i.e. free parameters, low, as our data set is rather small (Table II).

As we have a very low amount of data for a deep-learning problem, we chose also to design our network rather small (Tab. I). Additionally, the data were augmented on-the-fly using the methods: randomized rotation in a range of up to $4°$ as well as scaling up to $\pm 4 \%$, translation in x / y up to 10 / 1%, and flip in x. As mentioned above, we would like to reduce the level of augmentation in the direction of y on the chip border regions, thus we decrease the amount of translation in y and omit the flip in y. Finally, the data were contrast-normalized before being processed by the framework.

The classes in the problem are very imbalanced (ratio is 92.2% : 3.7% : 4.1%, Tab. II), hence they need to be balanced for the CNN. For this, we first try to weight the loss function higher when a class is less presented in the data set, but we found that this approach gives suboptimal accuracies for such a high imbalance. Hence, we tried a different approach and found that simply duplicating the samples for the underrepresented classes while applying data augmentation yields good results. The on-the-fly data augmentation ensures, despite several images are identical in the data set, that the images appear differently for the CNN after the augmentation. Therefore, such an approach produces a good amount of image variations on the fly from one source image [23].

B. Evaluation of the visual attention model

At first, we will verify if the attention model shows plausible human behavior. As we do not posses human eye tracking data for this task, and as there is, according to our literature search, no eye tracking data available for workers performing fault recognition in the semiconductor industry, we can only analyze if the neuronal activities and eye movements of the model are reasonable under a task like this and are in line with the general

Figure 5: Precisions for the region of interest selection of the attention model. The precision is measured in pixels via the distance between the true center to the determined one.

literature about visual search [41, 45, 49]. For this, we first illustrated the model's behavior on a representative example, as already shown during the model's description. This serves as a qualitative analysis. Afterwards, we quantify the results, i.e. the correctness of eye movements, which connotes here how many streets the attention model finds and how precise are the selected coordinates.

1) Correctness of the found streets: At first, we quantify the number of streets found. The model of visual attention finds 98.49% of the existing $32\,768$ streets in our data set correctly, whereby the false positives were 29 of these. The localization precision per each wafer is given in Tab. III. We defined a street as not found if the extracted center coordinates are not on the chip sides ($0.3 < x, y < 0.7$) or if the ROI borders are outside the image borders. The data set consists of $8\,192$ processable chips, and thus has a total of $32\,768$ street segments, based on the assumption that each chip has four street segments. The model's correctness of 98.49% is within the upper range of the shown accuracies in different tasks by previous literature. The first versions published in 2005 have achieved an accuracy of 50% [1, 41], whereas later ones have reached in a rather easy task with 3 objects 96.2% [36]. More recent versions have accomplished 92% in a larger and more complex real-world application with 100 objects [2, 39], and 85.4% in a different virtual reality application [40]. Therefore,

	Good	Bad
Good	0.96	0.04
Bad	0.12	0.88

Classified

(a) CNN with attention

	Good	Bad
Good	0.95	0.05
Bad	0.33	0.67

Classified

(b) CNN without attention

Figure 6: Normalized confusion matrices depicting for a given chip (or street) on the y-axis how the system classified this chip (x-axis). The classification rates are given as percentages. **a)** Confusion matrix of the CNN with the visual attention model beforehand. **b)** Confusion matrix of the CNN without visual attention.

Approach	Accuracy [%]	Fault detection accuracy [%]
Baseline		
KNN	70.52 ± 1.21	55.00
SVM	78.33 ± 1.34	64.40
MLP	65.26 ± 5.85	63.40
CNN [14]	75.24 ± 1.88	57.20
CNN [11]	78.89 ± 3.03	62.20
CNN [15]	78.47 ± 0.69	60.20
Our CNN	80.83 ± 2.38	67.40
Attention-based		
Attention + KNN	76.91 ± 1.49	57.80
Attention + SVM	85.43 ± 0.91	74.60
Attention + MLP	79.85 ± 3.52	73.25
Attention + CNN [14]	88.66 ± 0.88	80.40
Attention + CNN [11]	87.63 ± 1.20	76.80
Attention + CNN [15]	84.86 ± 1.82	70.60
Attention + Our CNN	$\mathbf{91.91 \pm 0.57}$	**87.80**

Table IV: Street and chip classification results. Mean accuracies and standard deviations for (i) baseline solutions, and (ii) attention approaches based on our model. The CNNs [14], [11], [15] denote the state-of-the-art in the wafer domain.

we conclude from this comparison that the model shows a very good performance to select the streets correctly via eye movements.

2) Precision of the region of interest extraction: To ensure a high-quality region of interest (ROI) extraction, we measured the centers' accuracies of the resulting street segments, whereas the street's center should be situated in the center of the image. These results were obtained by manually annotating the street centers of 200 randomly selected samples and measuring the pixel-based distance to the real image center. The attention model achieves in average a deviation in X and Y of 0.34 and 0.57 pixels respectively (Fig. 5), with corresponding standard deviations of 3.16 and 3.65 pixels. Humans typically make an eye movement within the vicinity of an object, and another eye movement(s) to fine-center the gaze. The second operation is not part of the model, but despite this, the model finds the ROIs quite accurately with mean deviations within the range of a few pixels. Hence, the model is probably more precise than as expected from human psychological findings.

In summary, the model shows reasonable behavior as expected during a visual search task like this, and achieves a high accuracy to localize the streets correctly.

C. Performance of the CNN for street classification

We compare here the performance of the full system with a system where visual attention was disabled.

1) Performance of the CNN with visual attention: The full system achieves an accuracy to classify streets correctly of 91.91 % (Tab. IV, mean accuracy over all classes). This and all other following values were measured over 5 runs. Additionally, we obtain a confusion matrix (Fig. 6a). We are especially interested in the precision to detect faults on the wafer, and this accuracy is observable from the confusion matrix, showing faults are detected correctly with 87.80 % (also listen in Tab. IV).

2) Performance without visual attention: To highlight the benefits of visual attention, we benchmark a system where the attention model was removed from the pipeline (Tab. IV), and add for comparision also other methods from the domain of wafer dicing as baseline. The CNN in this system uses directly whole chip images as input. This CNN reaches a mean accuracy of 80.83 % (Tab. IV). This accuracy does not look so bad, however, the accuracy to detect a fault correctly is only 67.40 %. (Fig. 6b). This value is much lower showing the effect of the visual attention model. In fact, the overall error rates increase without attention from 8 % to 19 %, and for the faulty chips from 12 % to even 33 %. These values show that especially the detection of faulty chips benefits from visual attention.

Next, we evaluated our system in the context of related as well as DL-based systems in the domain of automated visual inspection for semiconductor manufacturing and wafer dicing (Tab. IV). While as of the current state-of-the-art automated visual inspection systems often ML-based approaches are being deployed, i.e KNN-, SVM-, and MLP-based classifiers, we include those approaches in our evaluation to provide a more comprehensive comparison. Besides them, the newest DNNs are tested [11, 14, 15]. This does not include DNNs that focus on e.g. process monitoring over time due the required time component [12], and contributions which do not provide sufficient details for a reimplementation [12, 13].

Our evaluation shows that the proposed visual attention incorporating system constitutes a notable improvement over existing solutions, in which single CNN based systems [11, 14, 15] constitute the current state-of-the-art. While the CNNs of [11, 14] are the current state-of-the-art and reported in their publication accuracies of 98.2 % and 96.2 % on their problems respectively, they reach here merely 75.24 % and 78.89 %. This implies for us our problem could be more challenging. A look at their data material confirms this, their data is more homogeneous and with much larger and thus easier

Approach	Without attention		With attention		Improvement of	
	Accuracy [%]	Fault detection accuracy [%]	Accuracy [%]	Fault detection accuracy [%]	Accuracy	Fault detection accuracy
DenseNet121 [10]	77.47 ± 1.57	56.80 ± 2.95	84.86 ± 1.38	71.00 ± 2.12	7.39	14.20
DenseNet121* [10]	80.47 ± 0.97	62.80 ± 2.49	84.33 ± 1.40	69.20 ± 2.95	3.86	6.40
InceptionV3 [10]	77.18 ± 1.35	57.20 ± 3.03	83.08 ± 1.57	66.80 ± 3.49	5.90	9.60
InceptionV3* [10]	76.27 ± 1.76	54.40 ± 3.97	82.42 ± 2.68	65.60 ± 5.32	6.15	11.20
MobileNetV2 [10]	71.08 ± 1.01	44.40 ± 2.51	83.83 ± 1.04	68.40 ± 2.07	12.76	24.00
MobileNetV2* [10]	76.99 ± 1.36	56.20 ± 3.03	84.19 ± 1.10	69.00 ± 2.24	7.21	12.80
ResNet50 [10]	75.95 ± 1.80	55.60 ± 4.77	84.04 ± 1.37	69.00 ± 3.00	8.09	13.40
ResNet50* [10]	79.01 ± 2.41	59.80 ± 5.22	84.77 ± 0.86	70.40 ± 1.82	5.76	10.60
VGG10 [10]	83.19 ± 1.64	71.00 ± 4.69	86.75 ± 2.08	74.60 ± 4.45	3.57	3.60
VGG13 [10]	82.68 ± 2.24	69.80 ± 5.07	86.30 ± 1.94	73.80 ± 3.96	3.63	4.00
Xception [10]	77.27 ± 0.52	57.40 ± 0.89	83.99 ± 0.90	68.60 ± 2.07	6.71	11.20
Xception* [10]	80.52 ± 0.74	62.80 ± 1.79	85.18 ± 0.88	70.60 ± 1.82	4.66	7.80
Our CNN	80.83 ± 2.38	67.40 ± 7.57	$\mathbf{91.91 \pm 0.57}$	$\mathbf{87.80 \pm 1.92}$	11.08	20.4

Table V: Performance evaluation of classical deep neural networks from computer vision. Tested once as standard solutions without attention, and once when combined with our attention approach. The last columns show the improvement through attention. *Transfer learning, pretrained on ImageNet.

to detect faults. In summary, we conclude that the proposed system is, at least from published approaches, currently the best system for this problem.

3) Standard deep networks and benefits of attention: In the next investigation, we benchmarked different standard DL-based approaches to (i) evaluate how they perform in comparison and thus could be used as alternative, and (ii) how much visual attention benefits them. While conventional DNNs are broadly applied to different real-world tasks, yet, they were also mainly developed for the recognition of real-world objects (e.g., ImageNet). Thus, the question arises how they perform in our use case with very different test data. Table V illustrates the resulting accuracies for different DNN-based models and architectures, showing our solution surpasses them too.

We found that all networks show a performance boost by visual attention, hence all networks benefit from visual attention. This illustrates that attention is a crucial principle and promotes its idea of zooming in. Surprisingly, the standard networks do not cope with attention very well, seen as the accuracies for the streets are lower than with our customized network. That implies the standard deep networks profit less from attention. Furthermore, transfer learning [10] benefits even less from attention. We suppose that is because in the original data set (ImageNet), the objects are not shown in high resolution or large sizes like the faults in our street regions, hence the image material does not transfer well enough to our use case. Hence, our conclusion here is that visual attention is able to show its strengths much better if it is not mixed with transfer learning. This is reasonable since the data material is different.

Therefore, we conclude that our proposed system improves and outperforms, as evaluated in comparison, current state-of-the-art solutions in the wafer dicing domain. Other standard deep neural network approaches show also lower accuracy percentages, and we evaluate generally the benefits of visual attention.

Figure 7: Street and chip classification ground truth visualized for good (•), anomaly (•), and faulty (•) streets and chips.

D. Performance of the whole system

Finally, the class of a chip is calculated from the classification of its four street regions. The full system reaches the following accuracy to classify the chips correctly: $91\,\%$. This rate would be the important one for a final production system. The result of the classified streets and chip error classes can then be illustrated, here as ground truth, as in Fig. 7. It comprises the street and chip classification test results for good (•), anomaly (•) and faulty (•) streets and chips according to the used addressing scheme of the wafer. The shown error classes can then continue to differ in their respective defect pattern to be further assessed by an inspector.

IV. Discussion about how to couple a visual attention model with a CNN

Several options exist how to couple a visual attention model with a CNN as delineated in the introduction of this work, e.g. [18, 20]–[24]. The classical approach is that attention selects a spatial region (region of interest, ROI [17]), which is subsequently passed to a classifier for later recognition. This idea came up at first with the saliency models in the late 1990s, which select a ROI for a later classifier [17]. The classifier can of course today also be a CNN. Variations of this approach are the glimpse approach from Google [18], where a more modern neural network selects regions very swiftly (glimpses), which are then fed into a CNN.

Alternatives to the ROI approach would be other developments like proto-objects, where regions are formed like an object [50]. They eliminate the drawback of the ROI approach that the region has always the shape of a rectangle, and hence may not fit a more naturally shaped object.

Another alternative would be that visual attention modulates the neuronal response (amplify or suppress), which is closer to the processing in the brain [32]. This approach starts to be utilized in the deep learning community to a small degree (e.g. [23]). The idea is to feed all area into the CNN, but enhance neuronal responses in the attended region, and suppress all neurons in other places. The approach has the advantage that it can shape the region more clearly. Additionally, the amplification and suppression is closer to the neuronal processing of the brain ([32, 33], original data e.g. [51]), because attention in the brain operates multiplicatively such as multiplicatively enhancing or suppressing neurons.

We consider also to utilize this novel modulation approach, but finally decided against it:

i) The modulation approach requires to feed entire images and not only the ROIs into the CNN. As the images are relatively large in our application (up to $1\,000$ pixels across), this would result in a large CNN and hence in a reduced processing speed, as well as, in a lot of free parameters and overfitting (we have only a few data here). On the other hand, the street-showing ROIs in our case have a resolution of only 400×60 pixels, which is many times smaller than the entire chip image.

ii) The advantage of the newer approach is that the region can be shaped more clearly and flexibly. However, this big gain is in our wafer application of not much benefit, as the most structures in our image material are anyway rectangles.

iii) Despite the approach's closer similarity to the brain's functionality, the neuroscientific review literature also reveals that the effects of attention are much more complex than a simple 'enhance attended' or 'suppress all others' [32, 33]. As we aim for a strong biological plausibility, we would not like to ignore these effects lighthearted, and rather aim for a separate research publication investigating how a deep neuronal network can be integrated into an attention model or vice versa in a biologically plausible way.

iv) Finally, we have a complicated task-instruction here. The workers' instruction is to check the cuts for potential faults, which is of course realized by looking at the cuts (i.e. streets). Hence, we have applied attention to the streets here. However, not a street alone need to be checked for cracks, but rather also the close region next to it, ranging from the close surrounding of the street over the street-chip border towards the inside of the chip. Therefore, the workers' task requires to transfer from the attended structure (streets) to another spatial area. As it is not really known how such task instructions are realized in the brain, we do not know how to implement this in the attention model. The region of interest approach instead solves this problem out-of-the-box.

Therefore, we evaluate it is better to use the region of interest approach, which solves the problem nicely and has advantages in our case.

V. Conclusion and outlook

In this contribution, we propose a novel system for automated visual fault detection by combining a biologically-plausible model of visual attention with a deep neural network. The process of automated visual fault detection in the domain of semiconductor manufacturing and laser-based wafer dicing constitutes one particularly challenging application area, as defect patterns often range within a size of only a few pixels / µm. This problem is challenging for traditional convolutional neuronal networks, and it is getting more challenging due to the heterogeneity and imbalance of the image material. Visual attention is well suited for this problem, but not much used in the semiconductor industry yet, for which we created the first deep learning system with attention. Our benchmark shows that visual attention improves the mean accuracy from 81% to 92%, and the accuracy to detect faults correctly from 67% to 88%. Hence, the error rate especially for the faults drops from 33% to 12%. These rates outperform notable other state-of-the-art systems in the domain, as well as the power of standard deep learning systems.

This work utilizes a biologically-plausible model that is deeply rooted in neuroscience [1, 2, 33, 43] for the combination of visual attention with deep learning approaches. As of the current state of the art, it is often unclear how both principles from visual attention and learning-based approaches should be combined in an application, furthermore they often lack biological plausibility and therefore tend to be unreliable or outperformed (Sec. I).

Future projects can be built on top of this system and can be enhanced with it, for example in other domains. The zooming-in approach is certainly not limited to the inspection of wafer faults, many other mechanical engineering problems may exist where small faults have to be recognized in a large amount of data as well as also surely in domains outside mechanical engineering. Moreover, future work can cover improvement for series production, or an application-specific optimization of the imagery system. The imagery system is currently standard and can be certainly enhanced and optimized for the current application.

REFERENCES

[1] F. H. Hamker, "The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision," *Comput Vis Image Underst*, vol. 100, no. 1, pp. 64–106, 2005.

[2] F. Beuth, "Visual attention in primates and for machines - neuronal mechanisms," Doctoral dissertation, Technische Universität Chemnitz, Germany, 2019.

[3] S.-H. Huang and Y.-C. Pan, "Automated visual inspection in the semiconductor industry: A survey," *Computers in Industry*, vol. 66, pp. 1–10, 2015.

[4] A. Kumar, "Computer-Vision-Based Fabric Defect Detection: A Survey," *IEEE Trans Ind Electron*, vol. 55, no. 1, pp. 348–363, 2008.

[5] A. Hooper, J. Ehorn, M. Brand, and C. Bassett, "Review of wafer dicing techniques for via-middle process 3DI/TSV ultrathin silicon device wafers," in *Proc ECTC 2015*, 2015, pp. 1436–1446.

[6] K. Rahim and A. Mian, "A Review on Laser Processing in Electronic and MEMS Packaging," *J Electron Packag*, vol. 139, no. 3, p. 30801, 2017.

[7] T. Schlosser, F. Beuth, M. Friedrich, and D. Kowerko, "A novel visual fault detection and classification system for semiconductor manufacturing using stacked hybrid convolutional neural networks," in *Proc ETFA 2019*, 2019, pp. 1511–1514.

[8] Fei-Long Chen and Shu-Fan Liu, "A neural-network approach to recognize defect spatial pattern in semiconductor fabrication," *IEEE Trans Semicond Manuf*, vol. 13, no. 3, pp. 366–373, 2000.

[9] L. Xie, R. Huang, N. Gu, and Z. Cao, "A novel defect detection and identification method in optical inspection," *Neural Computing and Applications*, vol. 24, no. 7-8, pp. 1953–1962, 2014.

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

[11] S. Cheon, H. Lee, C. O. Kim, and S. H. Lee, "Convolutional Neural Network for Wafer Surface Defect Classification and the Detection of Unknown Defect Class," *IEEE Trans Semicond Manuf*, vol. 32, no. 2, pp. 163–170, 2019.

[12] K. B. Lee, S. Cheon, and C. O. Kim, "A Convolutional Neural Network for Fault Classification and Diagnosis in Semiconductor Manufacturing Processes," *IEEE Trans Semicond Manuf*, vol. 30, no. 2, pp. 135–142, 2017.

[13] K. B. Lee and C. O. Kim, "Recurrent feature-incorporated convolutional neural network for virtual metrology of the chemical mechanical planarization process," *J Intell Manuf*, pp. 1–14, 2018.

[14] T. Nakazawa and D. V. Kulkarni, "Wafer map defect pattern classification and image retrieval using convolutional neural network," *IEEE Trans Semicond Manuf*, vol. 31, no. 2, pp. 309–314, 2018.

[15] J. O'Leary, K. Sawlani, and A. Mesbah, "Deep learning for classification of the chemical composition of particle defects on semiconductor wafers," *IEEE Trans Semicond Manuf*, vol. 33, no. 1, pp. 72–85, 2020.

[16] M. Carrasco, "Visual attention: the past 25 years." *Vision Research*, vol. 51, no. 13, pp. 1484–1525, 2011.

[17] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-based Visual Attention for Rapid Scene Analysis," *IEEE Trans Pattern Anal Mach Intell*, vol. 20, no. 11, pp. 1254–1259, 1998.

[18] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," in *arXiv preprint arXiv:1412.7755*, 2014.

[19] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 135.

[20] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-Aware Face Hallucination via Deep Reinforcement Learning," in *Proc CVPR 2017*, 2017, pp. 690–698.

[21] H. Wang, S. Tang, Y. Zhang, T. Mei, Y. Zhuang, and F. Wu, "Learning Deep Contextual Attention Network for Narrative Photo Stream Captioning," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. New York, New York, USA: ACM Press, 2017, pp. 271–279.

[22] M. Stollenga and J. Masci, "Deep Networks with Internal Selective Attention through Feedback Connections," in *Advances in Neural Information Processing Systems*, 2014, pp. 3545–3553.

[23] Z. Zhao and A. Kumar, "Improving periocular recognition by explicit attention to critical regions in deep neural network," *IEEE Trans Inf Forensics Security*, vol. 13, no. 12, pp. 2937–2952, 2018.

[24] Y. Cai, D. Du, L. Zhang, L. Wen, W. Wang, Y. Wu, and S. Lyu, "Guided attention network for object detection and counting on drones," *arXiv preprint arXiv:1909.11307*, 2019.

[25] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc CVPR 2019*, 2019, pp. 3146–3154.

[26] B. Chen and W. Deng, "Hybrid-attention based decoupled metric learning for zero-shot image retrieval," in *Proc CVPR 2019*, 2019, pp. 2750–2759.

[27] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang, "Visual attention consistency under image transforms for multi-label image classification," in *Proc CVPR 2019*, 2019, pp. 729–739.

[28] E. K. Miller and T. J. Buschman, "Cortical circuits for the control of attention." *Curr Opin Neurobiol*, vol. 23, no. 2, pp. 216–222, 2013.

[29] J. K. Tsotsos, Y. Liu, J. C. Martinez-Trujillo, M. Pomplun, E. Simine, and K. Zhou, "Attending to visual motion," *Comput Vis Image Underst*, vol. 100, pp. 3–40, 2005.

[30] L. G. Ungerleider, T. W. Galkin, R. Desimone, and R. Gattass, "Cortical connections of area V4 in the macaque." *Cerebral Cortex*, vol. 18, no. 3, pp. 477–499, 2008.

[31] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex." *Cerebral Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

[32] J. H. Reynolds and D. J. Heeger, "The normalization model of attention." *Neuron*, vol. 61, no. 2, pp. 168–185, 2009.

[33] F. Beuth and F. H. Hamker, "A mechanistic cortical microcircuit of attention for amplification, normalization and suppression." *Vision Research*, vol. 116, no. Part B, pp. 241–257, 2015.

[34] A. Thielscher, M. Kölle, H. Neumann, M. Spitzer, and G. Grön, "Texture segmentation in human perception: a combined modeling and fMRI study." *Neuroscience*, vol. 151, no. 3, pp. 730–6, 2008.

[35] F. T. Qiu, T. Sugihara, and R. von der Heydt, "Figure-ground mechanisms provide structure for selective attention." *Nature Neuroscience*, vol. 10, no. 11, pp. 1492–1499, 2007.

[36] M. Antonelli, A. Gibaldi, F. Beuth, A. J. Duran, A. Canessa, M. Chessa, F. H. Hamker, E. Chinellato, and S. P. Sabatini, "A hierarchical system for a distributed representation of the peripersonal space of a humanoid robot," *IEEE Trans Auton Mental Develop*, vol. 6, no. 4, pp. 259–273, 2014.

[37] A. Thielscher and H. Neumann, "Globally consistent depth sorting of overlapping 2d surfaces in a model using local recurrent interactions," *Biological Cybernetics*, vol. 98, no. 4, pp. 305–337, 2008.

[38] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: a Bayesian inference theory of attention." *Vision Research*, vol. 50, no. 22, pp. 2233–2247, 2010.

[39] F. Beuth and F. H. Hamker, "Attention as cognitive, holistic control of the visual system," in *Proc NCNC 2015*, T. Villmann and F.-M. Schleif, Eds., 2015, pp. 133–140.

[40] A. Jamalian, F. Beuth, and F. H. Hamker, "The performance of a biologically plausible model of visual attention to localize objects in a virtual reality," in *Proc ICANN 2016*, 2016, pp. 447–454.

[41] F. H. Hamker, "The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement." *Cerebral Cortex*, vol. 15, no. 4, pp. 431–447, 2004.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint:1409.1556*, 2014.

[43] M. Zirnsak, F. Beuth, and F. H. Hamker, "Split of spatial attention as predicted by a systems-level model of visual attention." *Eur J Neurosci*, vol. 33, no. 11, pp. 2035–2045, 2011.

[44] N. K. Logothetis, J. Pauls, and T. Poggiot, "Shape representation in the inferior temporal cortex of monkeys," *Current Biology*, vol. 5, no. 5, pp. 552–563, 1995.

[45] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychon Bull Rev*, vol. 1, no. 2, pp. 202–238, 1994.

[46] P. Pouget, I. Stepniewska, E. a. Crowder, M. W. Leslie, E. E. Emeric, M. J. Nelson, and J. D. Schall, "Visual and motor connectivity and the distribution of calcium-binding proteins in macaque frontal eye field: implications for saccade target selection." *Frontiers in Neuroanatomy*, vol. 3, p. 2, 2009.

[47] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, D. B. Rosen *et al.*, "Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps," *IEEE Trans Neural Netw*, vol. 3, no. 5, pp. 698–713, 1992.

[48] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[49] J. M. Wolfe, "Visual search," *Current biology*, vol. 20, no. 8, pp. R346–R349, 2010.

[50] D. Walther and C. Koch, "Modeling attention to salient proto-objects." *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[51] J. H. Reynolds, L. Chelazzi, and R. Desimone, "Competitive mechanisms subserve attention in macaque areas V2 and V4." *Journal of Neuroscience*, vol. 19, no. 5, pp. 1736–53, mar 1999.

Unit	Layer	Type	Output shape	Kernel size	Stride
conv1	conv1_1	conv	$92 \times 92 \times 32$	5×5	1
	conv1_2	conv	$90 \times 90 \times 48$	3×3	1
	pool1	max pool	$30 \times 30 \times 48$	3×3	3
	dropout1	dropout	$30 \times 30 \times 48$	/	/
conv2	conv2_1	conv	$28 \times 28 \times 64$	3×3	1
	conv2_2	conv	$26 \times 26 \times 96$	3×3	1
	pool2	max pool	$13 \times 13 \times 96$	2×2	2
	dropout2	dropout	$13 \times 13 \times 96$	/	/
conv3	conv3_1	conv	$11 \times 11 \times 144$	3×3	1
	conv3_2	conv	$9 \times 9 \times 192$	3×3	1
	pool3	max pool	$4 \times 4 \times 192$	2×2	2
	dropout3	dropout	$4 \times 4 \times 192$	/	/
fully conn	dense1	fully conn	192	/	/
	dropout4	dropout	192	/	/
	dense2	fully conn	2	/	/

Table VI: Layer configuration of the convolutional neuronal network for chip classification.

VI. APPENDIX

In the first part of the appendix, we will list the CNN without visual attention. It utilizes as input directly whole chip images, and classifies them into good and faulty chips. In the second part, we provide a wafer overview and the street and chip groundtruth for each wafer.

A. CNN without visual attention

When we compare our system to a system without attention, we remove the attention model. Hence, the deep neural network has to process whole chips, resulting in a new CNN (Tab. VI). We designed for a fair comparison both networks as similar as possible. However, as chips are squared and not rectangle-shaped anymore, and as the trick with reduced spatial pooling in one dimension is no longer necessary (Sec. II-E), the last stage contains a normal pooling stage and the pooling size has to be changed to 3×3 (Tab. VI). The squared CNN's input size was chosen to hold the same amount of input pixels, at least roughly. The street CNN has an input shape of $60 \times 192 = 11520$ pixels. This result in a squared size of $\sqrt{11520} = 107$ pixels. We round it to the next multiple of 32, to make it feasible for the warp-size in the GPU, resulting in the finally chosen size of 96×96. Otherwise, the networks are the same. To verify that the size does not inflict any side effects, we run in addition a verification test with a bigger size of 192×192 and found no differences in the accuracy.

B. Insight into our data material

We received ten different wafers from a semiconductor manufacturer with different sizes, faults, materials, and imaging conditions. In the following, we have displayed all individual wafers. After analyzing the data, we created schematic overviews of each wafer to illustrate the distribution of faults by showing for each wafer which chips and streets are good and which have faults. Additionally, anomalies are marked that represents intact streets, but with an unknown visual event on them. The first wafer is shown in the main result section (Fig. 7), and the remaining ones in this appendix (Fig. 8 - 10) as the figures are relatively large in size and detail. These schematic diagrams illustrate the distribution of faults and give an impression, along with the wafers overviews, about the size and shape of each wafer for the reader. The wafer images were again scaled down and slightly modified to protect the intellectual property of the company.

Figure 8: Wafer overview, and chip plus street ground truth of wafers 2 - 4. The latter is visualized for good (•), anomaly (•) and faulty (•) streets. The notation of the figure is identically to Fig. 7. The wafer images were again scaled down to 500×500 pixels to protect the intellectual property.

Figure 9: Wafer overview and street ground truth of wafers 5 - 7. The figure is identically labeled to Fig. 8.

Figure 10: Wafer overview and street ground truth of wafers 8 - 10. The figure is identically labeled to Fig. 8. The wafer 10 is not available as a whole image, hence the wafer is not displayed here. We have only been provided single chip images.