Study on the Influence of Multiple Image Inputs of a Multi-View Fusion Neural Network Based on Grad-CAM and Masked Image Inputs

Stephan Tilgner, Daniel Wagner, Kathrin Kalischewski, Jan-Christoph Schmitz and Anton Kummert
School of Electrical, Information and Media Engineering
University of Wuppertal
42119 Wuppertal, Germany
{tilgner, dawagner, kalischewski, jaschmitz, kummert}@uni-wuppertal.de

***Abstract*—Neural network models are used successfully in many applications like traffic sign recognition in the automotive context, cancer detection in medicine engineering, machine monitoring in the manufacturing industry, et cetera. However, the decisions of a neural network model for a particular input sample in a classification task are mostly nontransparent. We propose techniques to determine which input image of a Multi-View Fusion Neural Network has the most influence on the prediction of the model for a particular image sample pair and which regions in the input images are important. In addition, a trained Multi-View Fusion Neural Network is studied regarding the question of influence. The results are convincing and show that the studied model learned similar concepts like a human.**

***Index Terms*—Influence Visualization, Multi-View Fusion Neural Network, Grad-CAM, Convolutional Neural Network**

I. Introduction

Since the decision of a neural network model for a particular class in a classification task is mostly nontransparent, many attempts are done to develop methods that should fill this gap. In the last years, several visual-based explanation methods have been proposed. Some approaches maximize one chosen neuron, which should be analyzed, by optimizing the input image [1]–[5] or rather an intermediate activation map [6]. The authors from [7] and [8] proposed sample-based techniques to study the concepts, like viewing angle or the occurrence of special textures, which are learned by the network. In contrast, the methods from [1], [9]–[15] highlight the regions of the input image which have the most influence on the decision of a neural network. The approach from [9] uses activation maps from a Convolutional Neural Network (CNN) to determine the regions of the input which have a great influence on the decision of the model. This approach, called Class Activation Mapping (CAM), is unfortunately limited to the class of Fully-Convolutional Networks (FCN), which means that the networks contain only convolutional layers and not a fully-connected layer part, which is often used as a classifier. Therefore, the authors of [10] and [11] extended the CAM approach by introducing a partial derivative of the score for a particular class w.r.t. the last activation map of the model as a weighting of the activation maps. The former one is called Grad-CAM and uses the first partial derivative or rather the gradient. The latter approach is called Grad-CAM++, which uses higher partial derivatives for determining the weighting. Using higher derivatives improves the localization of the high influence regions in cases where more than one class is present in the input image, according to [11]. Since the neural network model has to be differentiable to realize the backpropagation method in the training phase, any neural network which contains convolutional layers can be analyzed by the two approaches if the user has access to the partial derivatives of the model. The authors of [12] and [13] use masked input images, which are propagated through the neural network and the prediction scores are recorded and analyzed. The main idea is that masking out regions, that are important for the network's prediction, results in a drop in the prediction score. These approaches can even be used if the user has no access to the internals of the model, like the partial derivatives. They can be utilized for virtually any model which takes an image and outputs a classification score. The disadvantage compared to the methods based on CAM is that for one heatmap generation

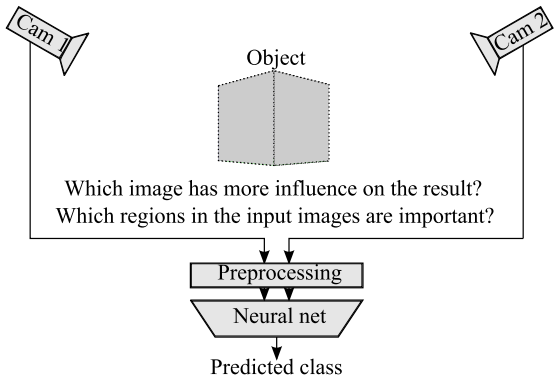


Fig. 1. Illustration of the problem statement. An image pair of the same scene is the input of an MVFNN, which predicts the class of the captured object. Important questions are: 1) Which camera image has more influence on the prediction result of the model? 2) Which regions in the input images are significant?

This research was supported by the European Regional Development Fund (Grant No: EFRE-0400216).

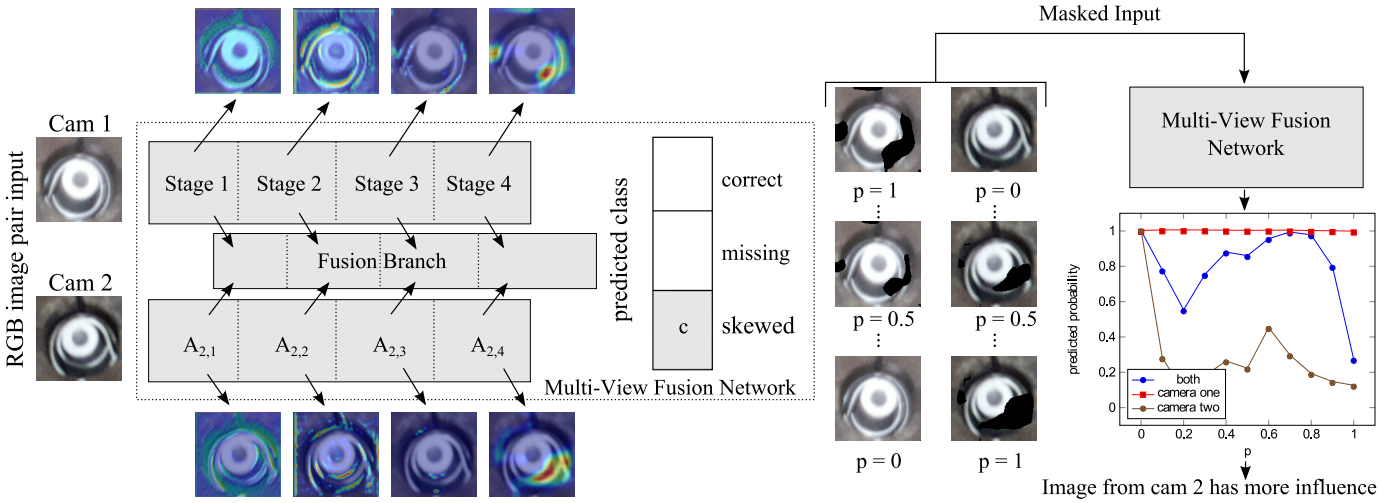


Fig. 2. Overview of the masked input approach. The Grad-CAM method is used to determine high influence regions of the two inputs separately at the last convolutional layers of the four stages of the outer branches of the MVFNN. The identified high influence regions from the last stages are employed to define binary masks which are parameterized by p. The variable p describes the proportion of the high influence region, which is set to zero in the mask. The impact of the inputs is studied by feeding the MVFNN with the masked inputs by varying p and analyzing the predicted probability curves for a particular class c.

many masked input images have to be processed by the model.

The proposed method in [12] uses square regions or rather bounding boxes as mask shapes. The final region of high influence is determined by agglomerative clustering. In [13] a method called Randomized Input Sampling for Explanation of Black-box Models (RISE) is proposed for determining the high influence regions. The main idea of this method is to occlude the input by many randomly chosen masks without constraints on the shape and to probe the masked images independently. Then the random masks are weighted by the score outcomes of the neural network, which is fed by the masked images. In the final step, the weighted masks are summed up for the heatmap of influence. However, all addressed methods assume only one input image. The recently proposed Multi-View Fusion Neural Network (MVFNN) [16] uses more than one input image of the same scene from different views for classification. The MVFNN model is used by the authors of [16] to predict the state of each mold cavity of an industrial mold injection machine before the injection process to reduce scrap. Here too the question arises of the regions of high influence. Furthermore, the question comes up which input has more influence on the prediction outcome.

This paper focuses on the question of determining the influence regions of multiple inputs of MVFNNs and to quantify which input has the most influence on the prediction score of a particular class. Figure 1 illustrates this problem statement.

II. Method

The main idea of the masked input method is to use the Grad-CAM approach to identify a prior of high influence regions of the inputs separately and to determine the influence of each input on the result by masking the inputs by a proportion of these regions. The process from generating the heatmaps of high influence regions by Grad-CAM to the analysis of these regions on the absolute influence on the model prediction is illustrated in Fig. 2. The studied MVFNN from [16] has a RGB image pair (I_1, I_2), with $I_i \in [0,1]^{u \times v \times 3}$, from the same mold cavity as input. The Grad-CAM heatmaps (cf. [10])

$$L^c_{i,j} = ReLU\left(\sum_{k=1}^{K}\left[\sum_{w=1}^{W}\sum_{h=1}^{H}\left(\frac{\partial y^c}{\partial A^k_{i,j}}\right)_{w,h}\right]A^k_{i,j}\right) \in \mathbb{R}_+^{W \times H} \tag{1}$$

are calculated at the four stages of the outer branches of the MVFNN for a particular class c, where $A^k_{i,j}$ is the k-th activation map for the i-th input of the last convolution layer at the j-th stage. Furthermore, W and H are the width and height of the activation map $A^k_{i,j}$ and y^c is the prediction score of class c. The $ReLU(\cdot)$ function is defined as

$$ReLU(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{2}$$

and is applied element-wise if the input is not a scalar.

The heatmaps $L^c_{i,j}$ are resized or rather interpolated to $\hat{L}^c_{i,j} \in \mathbb{R}_+^{u \times v}$ so that they fit the width and height of the input images. The resized heatmaps $\hat{L}^c_{i,4}$ of the last stage are used to generate the occlusion masks by using a threshold of $\frac{1}{4} \times \max(\hat{L}^c_{i,4})$. All positions of $\hat{L}^c_{i,4}$ with values higher than the threshold are determined and the corresponding values are sorted in descending order. A proportion $p \in [0,1]$ of the ordered value-position pair list is used to define the 0 entries in the occlusion mask $M^c_{i,p} \in \{0,1\}^{u \times v}$, where i is again the index of the input.

The occlusion mask $M^c_{i,p}$ is element-wise and channel-wise multiplied with the input image I_i. Three accuracy curves $a_l(p)$ are determined by feeding the model with the masked inputs $(I_1 \odot M^c_{1,1-p}, I_2 \odot M^c_{2,p})$ or rather $(I_1 \odot M^c_{1,p}, I_2)$ and $(I_1, I_2 \odot M^c_{2,p})$ for $p \in [0,1]$, where $\odot$ is the element-wise and channel-wise multiplication operator. The first curve (a_1)

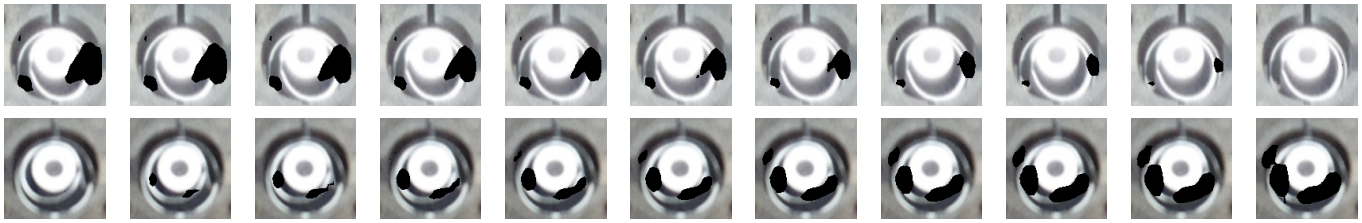
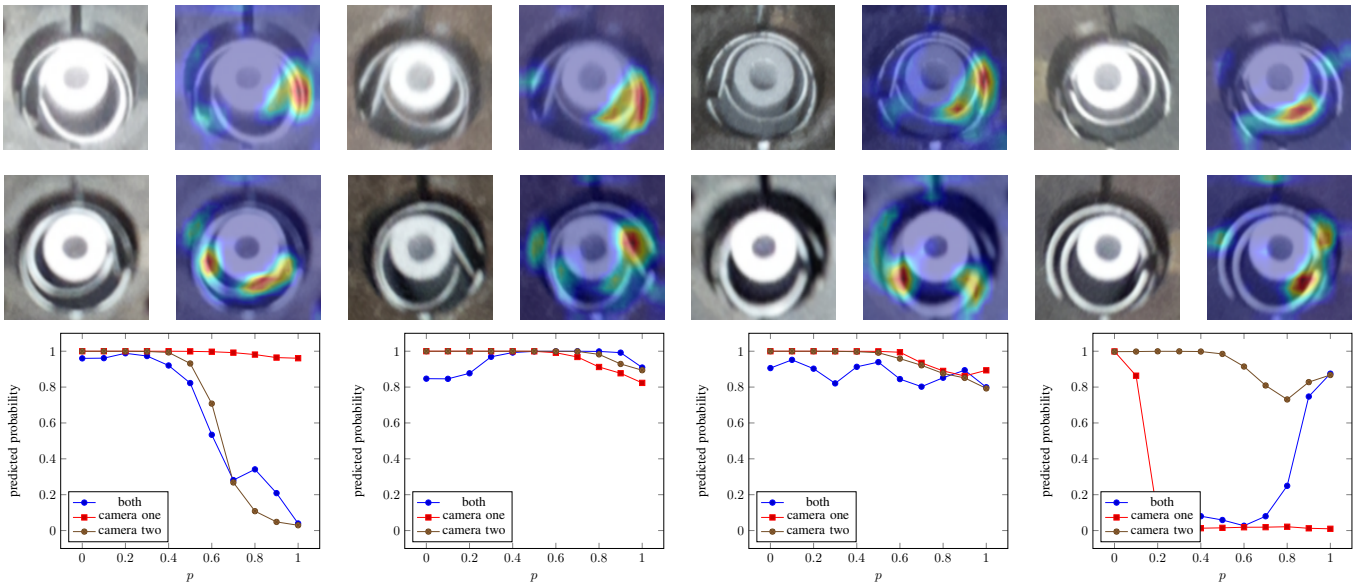


Fig. 3. Example of masked input pairs that are used to generate the accuracy curves of type one. One particular input pair per column is shown. In the first row examples of masked inputs, $I_1 \odot M^c_{1,1-p}$, of image one are shown. In the second row masked inputs, $I_2 \odot M^c_{2,p}$, of image two are illustrated, respectively. Both image rows are generated with $p = \{0, \frac{1}{10}, \frac{2}{10}, \ldots, 1\}$ in increasing order.



(a) $m = 0.346$ (b) $m = -0.021$ (c) $m = 0.016$ (d) $m = -0.731$

Fig. 4. Four examples (a-d) of the masked input approach for prediction the class skewed. The first row shows the images from camera one of the class skewed and the corresponding heatmaps from the Grad-CAM approach superimposed with the images in alternating order. The second row illustrates the same for images from camera two of the same mold cavity. The associated accuracy curves from the masked input approach are shown in the third row.

describes the couplet influence of the inputs and the second (a_2) and third curve (a_3) the independent influence of input from camera one and camera two, respectively. An example of the inputs $(I_1 \odot M^c_{1,1-p}, I_2 \odot M^c_{2,p})$ with $p = \{0, \frac{1}{10}, \frac{2}{10}, \ldots, 1\}$ in increasing order is shown in Fig. 3.

The accuracy curves a_l are analyzed to determine which camera image has the most influence on the prediction of the trained MVFNN. The mean of the differences

$$m = \frac{1}{\#p} \sum_p [a_2(p) - a_3(p)] \in [-1, 1] \tag{3}$$

is an indicator for the question which image has more impact on the prediction. If m is negative with a high magnitude then the image from camera one has more impact on the prediction than the image from the same scene from camera two. Moreover, if m is positive with high magnitude, then the image from camera two has more influence on the prediction. If m is near zero, we assume an equal contribution of both camera images to the prediction. In other words, the sign of m describes which image has more impact and the magnitude quantifies how strong the influence is. We define equal impact if $|m| < \frac{2}{100}$.

III. Experiments

For the evaluation of the proposed method, one hundred randomly picked sample pairs from each of the three classes {correct, missing, skewed} are studied to analyze a trained MVFNN from [16]. Four results of the class skewed are shown in Fig. 4.

The first row shows the images from camera one and the corresponding heatmaps from the Grad-CAM approach superimposed with the images in alternating order. The second row illustrates the same for images from camera two of the same mold cavity. The associated accuracy curves and corresponding values of m (cf. (3)) are shown in the third row. It is evident from Fig. 4 that the trained MVFNN focuses on regions of the input images which should be focused by a human for classifying the state of the mold cavity into the three classes {correct, missing, skewed}. The regions where the inner bowl has a small distance to the outer bowl are

important for the trained MVFNN to predict the class skewed, which is also true for a human observer.

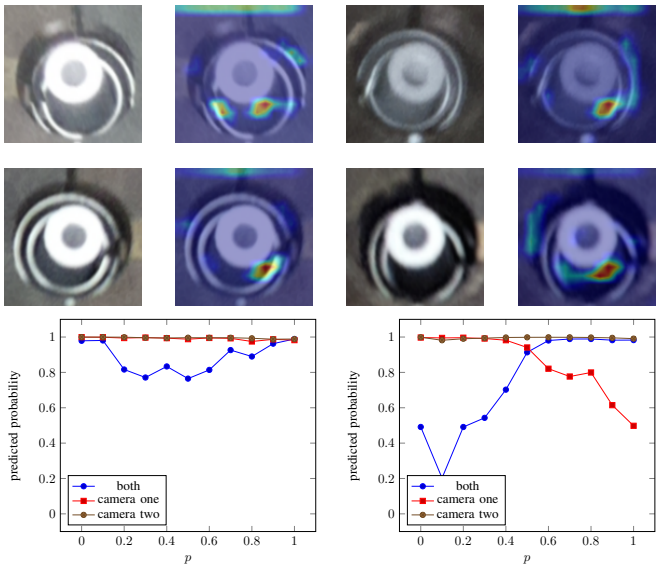
In the first example (Fig. 4(a)), the image from camera two has more impact on the prediction since the accuracy curves for camera two and the coupled curve for both cameras collapse if the proportion of the occlusion mask is raised and the accuracy curve for the masked image from camera one does not collapse. The mean of the differences in this example is $m = 0.346$. This is feasible for a human observer since in the image from camera two the inner bowl contacts the core.

In the second example (Fig. 4(b)), the image from camera one has more influence since $m = -0.021$. It is not as clear as in the first example since the accuracy curves for the masked image of camera one and two degenerate both with increasing the proportion of the occlusion mask which results in a low magnitude of m. However, the curve for the masked input from camera one degenerates faster than the curve for the masked input image from camera two. Moreover, the accuracy curve for both masked input images is lower at the beginning, where the input image from camera one is more occluded as at the end, where camera image two is occluded. Since the accuracy curves decrease to an absolute value of approximately 80% and the magnitude of m is low in contrast to the first example (Fig. 4(a)), the conclusion can be drawn that the regions of high influence in the second example (Fig. 4(b)) are not as important as in the first.

The mean of differences is $m = 0.016$ in the third example (Fig. 4(c)). Hence, the accuracy curves for the images of camera one and camera two are very close and the magnitude of m is below the threshold of $\frac{2}{100}$. Furthermore, the curve for both masked images oscillates. Therefore, the conclusion can be drawn that even though the image from camera two has slightly more impact, both input images are almost equally important. The image from camera one in the last example (Fig. 4(d)) has more influence on the prediction of the MVFNN because of the big gap between the accuracy curves. The resulting mean of differences is $m = -0.731$.

In Fig. 5, two examples of the class correct are shown. In example Fig. 5(a), both images have equal impact on the prediction of the trained MVFNN since the accuracy curves for the masked input image from camera one and camera two are close together which result in a mean of differences of $m = -0.004$. For a human observer, it is also true, both images have the same amount of information for predicting the sample pair as class correct. In example Fig. 5(b) the image from camera one has the most influence with $m = -0.139$. This is also true for the human viewer since the image from camera two contains a black shadow in spaces where the inner bowl is placed. The black shadow occurs because the camera view vector is not perpendicular to the mold cavity plane. Therefore, there is a blind spot and after an inverse perspective mapping in a prepossessing step, it results in a black shadow (cf. [16]).

Besides, an example of the class missing is shown in Fig. 6. The influence is equal since all accuracy curves are close-by one resulting in $m = -0.001$. This is representative of the hundred studied image pairs of the class missing. Indeed, one image is enough to predict the inner bowl as missing by a human observer.



(a) $m = -0.004$ (b) $m = -0.139$

Fig. 5. Two examples (a-b) of the class correct. The first row shows the images from camera one and the corresponding Grad-CAM heatmaps superimposed with the images in alternating order. The second row illustrates the same for images from camera two of the same mold cavity. The associated accuracy curves from the masked input approach are shown in the third row.

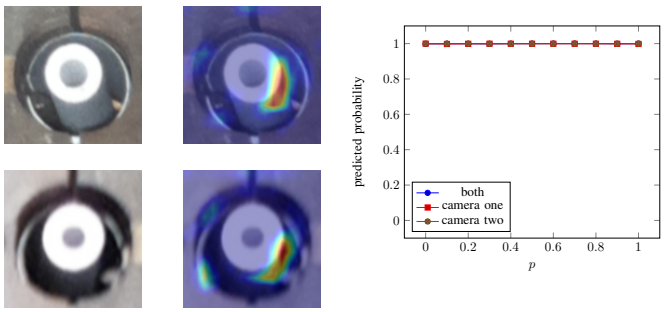


Fig. 6. An example of the class missing. Both input images have equal influence ($m = -0.001$) on the prediction of the MVFNN.

IV. Conclusion

We proposed an approach for determining the influence of input images of a trained Multi-View Fusion Neural Network (MVFNN) on the prediction of a particular class. In other words, we proposed a method that can tell the user which input image has more impact on the prediction. The masked input approach is transparent to the user and experiments show that this method produces reliable results. Moreover, experiments show that a well trained MVFNN for predicting the state of an industrial mold injection machine uses similar concepts as a human observer, like focusing on regions where the inner bowl has a small distance to the outer bowl for sample pairs of the class skewed.

REFERENCES

[1] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6034

[2] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 5188–5196.

[3] A. Nguyen, J. Yosinski, and J. Clune, "Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks," *arXiv preprint arXiv:1602.03616*, 2016.

[4] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.

[5] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug play generative networks: Conditional iterative generation of images in latent space," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3510–3520.

[6] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," University of Montreal, Technical Report, 2009.

[7] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 2668–2677. [Online]. Available: http://proceedings.mlr.press/v80/kim18d.html

[8] M. Aubry and B. C. Russell, "Understanding deep features with computer-generated imagery," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2875–2883.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2921–2929.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 618–626.

[11] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2018, pp. 839–847.

[12] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani, "Self-taught object localization with deep networks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–9.

[13] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *BMVC*, 2018.

[14] A. Binder, S. Bach, G. Montavon, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for deep neural network architectures," in *Information Science and Applications (ICISA) 2016*, K. J. Kim and N. Joukov, Eds. Singapore: Springer Singapore, 2016, pp. 913–922.

[15] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. [Online]. Available: https://openreview.net/forum?id=BJ5UeU9xx

[16] S. Tilgner, D. Wagner, K. Kalischewski, J. Velten, and A. Kummert, "Multi-view fusion neural network with application in the manufacturing industry," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2019, pp. 1–5.