ANOMALY DETECTION AND LOCALIZATION IN IMAGES USING GUIDED ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Anomaly detection and localization is a popular computer vision problem which involves detecting anomalous images and localizing anomalies within them. However, this task is challenging due to small sample size and pixel coverage of the anomaly in real-world scenarios. Previous works have a drawback of using anomalous images to compute a threshold during training to detect and localize anomalies. To tackle these issues, we propose AVAGA - the first end-to-end trainable convolutional adversarial variational autoencoder (CAVAE) framework using guided attention which localizes the anomaly with the help of attention maps. AVAGA detects an image as anomalous from the large pixel-wise difference between the input and reconstructed image. In an unsupervised setting, we propose a guided attention loss, where we encourage AVAGA to focus on all non-anomalous regions in the image without using any anomalous images during training. Furthermore, we also propose a selective gradient backpropagation technique for guided attention, which enhances the performance of anomaly localization while using only 2% anomalous images in a weakly supervised setting. AVAGA outperforms the state-of-the-art (SoTA) methods by 10% and 18% on localization (IoU) and 8% and 15% on classification accuracy in unsupervised and weakly supervised settings respectively on Mvtec Anomaly Detection (MvAD) dataset and by 11% and 22% on localization (IoU) and 10% and 19% on classification accuracy in unsupervised and weakly supervised settings respectively on the modified ShanghaiTech Campus (STC) dataset.

1 INTRODUCTION

With several breakthroughs of Deep Neural Networks (DNNs) outperforming humans in the field of image classification (He et al. (2016)), action recognition (Girdhar et al. (2019)), face recognition (Liu et al. (2017)), etc., one area where it has made significant progress is recognizing whether an image is homogeneous with its previously observed distribution or whether it belongs to a novel or anomalous distribution (Akcay et al. (2018)). To develop machine learning algorithms for such a setting can be challenging due to the lack of suitable data since images with anomalies are rarely available in real world scenarios as discussed by Bergmann et al. (2019). Previous works on anomaly detection (Benezeth et al. (2009), Böttger & Ulrich (2016), Steger (2001)) employ handcrafted features to detect anomalies, while Hasan et al. (2016) and Dimokranitou (2017) propose autoencoder based networks in such challenging settings. GAN based approaches (Ravanbakhsh et al. (2019), Zenati et al. (2018)) have also been proposed for this task. Wang et al. (2018), Tran & Yuan (2011) propose temporal anomaly localization while Cheng et al. (2013) propose patch based anomaly localization in videos. These approaches train their network with non-anomalous images / videos and use a thresholded pixel-wise difference between the input and reconstructed image to detect anomalous images and localize anomalies within them. However, their drawback is that the threshold needs to be computed by using anomalous images during training.

To solve this drawback, we propose AVAGA - a convolutional adversarial variational autoencoder network with guided attention to address anomaly detection and localization in two different training settings i.e. unsupervised and weakly supervised. In an unsupervised setting, given the limited sample size and the pixel coverage of anomaly in the image, we encourage the network to focus on all non-anomalous regions of the image such that the feature representation of the latent space encodes all the non-anomalous regions. Following Bergmann et al. (2019), we train our network in

an unsupervised setting comprising of only non-anomalous images. We denote non-anomalous as normal for the rest of our discussion. In the weakly supervised setting we introduce a classifier in our network and propose the idea of selective gradient backpropagation for guided attention, where we compute an attention map for only the images correctly predicted by the classifier to localize the anomaly better.

To the best of our knowledge, we are the first to propose an end-to-end trainable framework that guides the network in learning an attention map to localize the anomaly in both unsupervised and weakly supervised settings. As compared to the prior works, our proposed approach does not use any anomalous images during training to compute a threshold to detect and localize the anomaly. Our contributions are: (a) An end-to-end trainable convolutional adversarial variational autoencoder which comprises of a convolutional latent space to preserve the spatial relation between the input and reconstructed image. (b) A guided attention loss which is jointly optimized with adversarial reconstruction training to detect and localize the anomaly in an unsupervised setting. (c) A selective gradient backpropagation technique for guided attention, to control any incorrect attention map generated from the prediction of the classifier to localize the anomaly better. AVAGA outperforms SoTA methods on MvAD dataset (Bergmann et al. (2019) by 10% and 18% on localization and 8% and 15% on classification accuracy in unsupervised and weakly supervised settings respectively. The experiments on the modified STC dataset (Liu et al. (2018)) also outperform SoTA approaches by 11% and 22% on localization and 10% and 19% on classification accuracy in unsupervised and weakly supervised settings respectively.

2 Proposed Approach

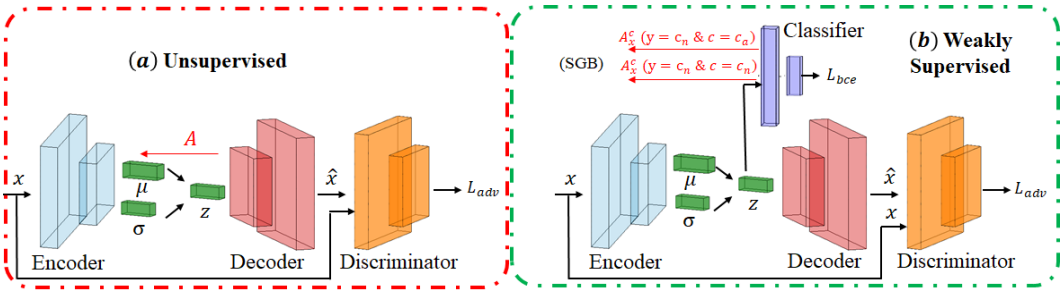

Figure 1: (a) Framework of AVAGA$_u$ where the attention map (A) is computed from the latent space z using Grad-CAM. (b) Illustration of AVAGA$_w$ with selective gradient backpropagation for guided attention (SGB) to compute the attention maps (A_x^c) from the classifier's prediction. The architectural details are presented in Table 9 of Appendix.

2.1 Unsupervised Approach: AVAGA$_u$

We discuss the idea of a CAVAE using guided attention (AVAGA$_u$) as shown in Figure. 1(a) where we encourage the network to learn a feature representation of the latent space z by training it using a reconstruction loss with adversarial learning (L_{adv}) on only normal images. Since attention maps obtained from feature maps illustrates the regions of the image responsible for specific activation of neurons in it (Zagoruyko & Komodakis (2016)), we propose a guided attention loss and use it as an additional supervision to guide the network to focus on all the normal regions of the image, such that the anomalous attention map localizes the anomaly during inference.

2.1.1 Convolutional Adversarial Variational Autoencoder (CAVAE)

Variational Autoencoder (VAE) (Kingma & Welling (2013)) is a generative model which is widely used for anomaly detection (Pawlowski et al. (2018)). The formulation for training a vanilla VAE is as follows:

$$L(x,\hat{x}) = L_R(x,\hat{x}) + KL(q_\phi(z|x)||p_\theta(z|x))$$

$$\text{where, } L_R(x,\hat{x}) = -\frac{1}{N}\sum_{i=1}^{N} x_i log(\hat{x}_i) + (1-x_i)log(1-\hat{x}_i) \quad (1)$$

Here, x is the input image, $\hat{x}$ is the reconstructed image and N is the total number of images. The latent space $p_\theta(z|x)$ is modeled using a simple prior $p(z)$ (standard Gaussian distribution) with the help of Kullback-Liebler (KL) divergence through $q_\phi(z|x)$. Since, the vanilla VAE results in blurry reconstruction (Larsen et al. (2015)), we use a discriminator ($D(.)$) to improve the stability of the training and generate a sharper reconstruction $\hat{x}$ using adversarial learning (Makhzani et al. (2015)) formulated as follows:

$$L_{adv} = -\frac{1}{N}\sum_{i=1}^{N} log(D(x_i)) + log(1 - D(\hat{x_i})) \tag{2}$$

Unlike traditional autoencoders (Gutoski et al. (2017), Bergmann et al. (2018)) where the latent space is vectorized, inspired from (Myronenko (2018)) we propose an end-to-end CAVAE to preserve the spatial relation between the input and the reconstructed image. We illustrate the effectiveness of using a convolutional latent space over vectorizing it in Sec. 5.

2.1.2 GUIDED ATTENTION

Along with detecting an image as anomalous, we also focus on spatially localizing the anomaly in the image. Most works (Schlegl et al. (2017), Vu et al. (2019), Akcay et al. (2018)) employ a thresholded pixel-wise difference between the reconstructed image and the input image to localize the anomaly in which the threshold is determined by using anomalous images during training. However, AVAGA$_u$ learns to localize the anomaly using an attention map reflected through an end-to-end training process without the need for any anomalous images.

We use the feature representation of the latent space z to compute the attention map. Attention map (A) is computed using Grad-CAM (Selvaraju et al. (2017)) and normalized using a Sigmoid operation such that $A_{i,j} \in [0, 1]$ to make it differentiable during the end-to-end training process.

Intuitively, the attention map obtained from the feature map focuses on certain regions of the image based on the activation of neurons and its respective importance (Zhou et al. (2016), Zagoruyko & Komodakis (2016)). Since our training set consists only of normal images, we intend to learn the feature representation of the entire image. We use this notion to propose a guided attention loss to provide extra supervision to the network and encourage it to generate an attention map that covers all the normal regions such that the feature representation of the latent space encodes all the normal regions. This guided attention loss is formulated as follows:

$$L_{attn} = \frac{1}{Z}\sum_{i,j}(1 - A_{i,j}) \tag{3}$$

Here, Z is the number of pixels in A. Using eq. 1, eq. 2 and eq. 3, we formulate our final objective function L_{final} as follows:

$$L_{final} = w_r(L_R(x, \hat{x})) + w_{kl}(KL(q_\phi(z|x)||p_\theta(z|x))) + w_{adv}(L_{adv}) + w_{attn}(L_{attn}) \tag{4}$$

The magnitude of each loss is balanced using scaling factors w_r, w_{kl}, w_{adv} and w_{attn} which is set as $1, 1, 1$ and $1e^{-2}$ respectively from validation.

During testing, we input image x_{test} into the AVAE, which reconstructs an image $\hat{x_{test}}$. The pixel-wise difference between $\hat{x_{test}}$ and x_{test} results in an anomalous score which detects x_{test} as an anomaly. Intuitively, if distribution of x_{test} is similar to the distribution of the learnt latent space, then the anomalous score is small. The attention map A_{test} is computed from the latent space using Grad-CAM and is inverted (**1** - A_{test}) to obtain an anomalous attention map which localizes the anomaly. Here, **1** refers to a unit matrix with same dimension as A_{test}.

2.2 WEAKLY-SUPERVISED APPROACH: AVAGA$_w$

AVAGA$_u$ can be further extended in a weakly supervised setting (AVAGA$_w$) where we explore the possibility of using few anomalous images during training to improve the performance of anomaly detection and localization. Based on previous works (Selvaraju et al. (2017), Oquab et al. (2015)), attention maps generated from a trained classifier have been used in weakly-supervised semantic segmentation tasks. Given the labels of the anomalous and normal images without the pixel-wise annotation of the anomaly during training, we modify AVAGA$_u$ by introducing a binary classifier at the output of the latent space as shown in Figure. 1(b) and train it using binary cross entropy loss

(L_{bce}) on normal and anomalous images along with training the CAVAE using adversarial reconstruction loss (eq. 1 + eq. 2). Since the attention map depends on the performance of the classifier (Li et al. (2018)), we propose a selective gradient backpropagation for guided attention (SGB) to compute an attention map based on the classifier's prediction to localize the anomaly better.

2.2.1 SELECTIVE GRADIENT BACKPROPAGATION FOR GUIDED ATTENTION

Given an image x and its corresponding label y, we define $c \in \{c_a, c_n\}$ as the prediction of the classifier, where c_a is anomalous and c_n is normal prediction respectively. From figure 1(b), we introduce a binary classifier in AVAGA$_w$ which is obtained by cloning the latent space z into a new tensor and flattening it to form a fully connected layer. The weights between z and its clone are shared. For the purpose of classification, we choose to separately vectorize the latent space which also enables the higher magnitude of gradient flow from the classifier's prediction to compute the attention map. (Selvaraju et al. (2017)).
In the pipeline of AVAGA$_w$ we train our latent space using normal images and propose an anomalous attention loss in which we focus on localizing the anomaly only in the anomalous images and prevent any anomalous attention map on the normal image. Using Grad-CAM we compute the anomalous attention map (A_x^c) from the anomalous prediction $c = c_a$ on image x when $y = c_n$. In addition to anomalous attention loss, we propose a normal attention loss where we focus on generating an attention map that covers all the normal regions of the image. The normal attention map is computed from the normal prediction $c = c_n$ on the image x when $y = c_n$. Thus, with the anomalous and normal attention loss, we encourage the network to focus on all the normal regions while preventing any anomalous attention on the normal image. The guided attention loss L_a in the weakly supervised setting is formulated as follows:

$$L_a = \begin{cases} \frac{1}{Z}\sum_{i,j}(A_x^c)_{i,j} & \text{if } c = c_a \text{ and } y = c_n \\ \frac{1}{Z}\sum_{i,j} 1 - (A_x^c)_{i,j} & \text{if } c = c_n \text{ and } y = c_n \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Since the attention map is computed by backpropagating the gradients from the classifier's prediction, any incorrect prediction would generate an undesired attention map. This would lead to the network learning to focus on erroneous regions of the image during training which we avoid using a technique termed as "selective gradient backpropagation for guided attention". Using label information and classifier's output, we compute L_a for only the images correctly classified by the classifier i.e. if $y = c$. From eq. 1, eq. 2 and eq. 5 based on SGB, we train our network using the following objective function:

$$L_{final} = w_r(L_R(x, \hat{x}) + w_{kl}(KL(q_\phi(z|x)||p_\theta(z|x))) + w_{adv}(L_{adv}) + w_c(L_{bce}) + w_a(L_a) \tag{6}$$

The magnitude of each of the losses is balanced using scaling factors w_r, w_{kl}, w_{adv}, w_c and w_a which is set as $1, 1, 1, 1e^{-3}$, and $1e^{-2}$ respectively from validation.
During testing, we input image x_{test} to AVAGA$_w$ which the classifier identifies as anomalous or normal. The anomalous attention map A_{test} is computed on x_{test} if $y = c_a$ using Grad-CAM which localizes the anomaly.

3 EXPERIMENTAL EVALUATION

Experiments are performed on MvAD and modified STC datasets to evaluate the performance of AVAGA$_u$ and AVAGA$_w$. We summarizes the datasets used for evaluation in Table 1 and discuss the implementation details in Sec. A.1 of Appendix. Based on the framework in Figure. 1(a), we use the convolution layers of ResNet-18 (He et al. (2016)) as our encoder pretrained from ImageNet (Russakovsky et al. (2015)) and fine-tune on the each category / scenes individually. Inspired from (Brock et al. (2018)), we propose to use the residual generator as our residual decoder by modifying it with a convolution layer interleaved between two upsampling (transpose convolution) layers to preserve local spatial information during reconstruction. The skip connection is added from the output of the upsampling layer to the output of the convolution layer to preserve high level feature information across upsampling layers. We use the discriminator of DC-GAN (Radford et al. (2015))

pretrained on Celeb-A dataset (Liu et al. (2015)) and finetune on our data as our discriminator. This network is termed as AVAGA-R and it's architectural details is presented in Table 9 of Appendix.
We illustrate the effectiveness of AVAGA$_u$ and AVAGA$_w$ over 1) AE L2 (Bergmann et al. (2018)) 2) AE SSIM (Bergmann et al. (2018)) 3) AnoGAN (Schlegl et al. (2017)) 4) CNN feature dictionary (Napoletano et al. (2018)) 5) Texture inspection (Böttger & Ulrich (2016)) 6) Variation model (Steger (2001)) based approaches. For fair comparisons with the baseline autoencoder and GAN based approaches, we employ the discriminator and generator of DCGAN pretrained on Celeb-A dataset as our encoder and decoder respectively, and use the same discriminator as discussed previously and train this network (AVAGA-D) using eq. 4 & eq. 6 and evaluate its performance for localization and classification accuracy. During our discussion we refer AVAGA-D$_u$ & AVAGA-R$_u$ jointly as AVAGA$_u$ in unsupervised and AVAGA-D$_w$ & AVAGA-R$_w$ as AVAGA$_w$ in weakly supervised setting respectively.
From Table 1, we observe that in the unsupervised setting, the network is trained only on the normal images. However in the weakly supervised setting, since none of the baseline methods provide information on the number of anomalous images they use to compute the threshold during training, we randomly choose 2% of anomalous images along with the complete set of normal image for training. Following Bergmann et al. (2019), we use the mean of accuracy of correctly classified anomalous images and normal images to evaluate the performance of anomaly detection and Intersection-over-Union (IoU) between the generated attention map and the ground truth segmentation mask to evaluate localization performance.

Table 1: Summary of the different datasets used to evaluate AVAGA$_u$ and AVAGA$_w$.

Dataset	MvAD	MvAD	modified STC	modified STC
Setting	unsupervised	weakly supervised	unsupervised	weakly supervised
# Categories / Scenes	15 categories	15 categories	13 scenes	13 scenes
# Train images	3629 normal	3664 normal & anomalous	244875 normal	246638 normal & anomalous
# Test mages	1725 anomalous	1690 anomalous	88167 anomalous	86404 anomalous
Loss function	eq. 4	eq. 6	eq. 4	eq. 6

Table 2: Comparison of IoU of AVAGA$_u$ and AVAGA$_w$ with state-of-the-art approaches on the MvAD dataset. The color of the highlighted number denotes the ranking performance, darker color indicates better performance.

Category	AE SSIM	AE L2	AnoGAN	CNN feature dictionary	Texture inspection	Variation model	AVAGA-D$_u$	AVAGA-R$_u$	AVAGA-D$_w$	AVAGA-R$_w$
Bottle	0.15	0.22	0.05	0.07	-	0.03	0.28	0.33	0.36	0.39
Hazelnut	0.00	0.41	0.02	0.00	-	-	0.42	0.47	0.58	0.79
Capsule	0.09	0.11	0.04	0.00	-	0.01	0.24	0.27	0.38	0.41
Metal Nut	0.01	0.26	0.00	0.13	-	0.19	0.38	0.45	0.46	0.46
Leather	0.71	0.67	0.34	0.74	0.98	-	0.75	0.79	0.80	0.84
Pill	0.07	0.25	0.17	0.00	-	0.13	0.31	0.38	0.44	0.53
Wood	0.36	0.29	0.14	0.47	0.51	-	0.55	0.59	0.61	0.66
Carpet	0.69	0.38	0.34	0.20	0.29	-	0.71	0.73	0.70	0.81
Tile	0.04	0.23	0.08	0.14	0.11	-	0.29	0.32	0.68	0.81
Grid	0.88	0.83	0.04	0.02	0.01	-	0.30	0.32	0.42	0.55
Cable	0.01	0.05	0.01	0.13	-	-	0.34	0.43	0.49	0.51
Transistor	0.01	0.22	0.08	0.03	-	-	0.28	0.34	0.38	0.45
Toothbrush	0.08	0.51	0.07	0.00	-	0.24	0.54	0.55	0.60	0.63
Screw	0.03	0.34	0.01	0.00	-	0.12	0.39	0.48	0.51	0.66
Zipper	0.10	0.13	0.01	0.00	-	-	0.18	0.25	0.29	0.31

4 Comparision with State-of-the-art

We compare AVAGA$_u$ and AVAGA$_w$ with baseline approaches on MvAD and modified STC datasets. We see from Table 2 that AVAGA$_u$ localizes the anomaly better compared to other baseline methods in the unsupervised setting on the MvAD dataset. Specifically, in 13 out of 15 categories, AVAGA-D$_u$ outperforms the baseline with the best performance in these categories with an improvement ranging from 1% to 21%. Although the major focus of our work is in localizing the anomaly

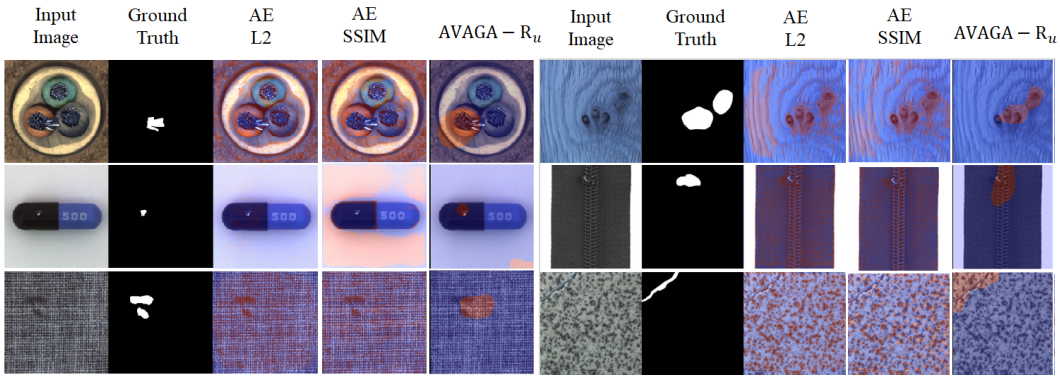

Figure 2: Qualitative results on the MvAD dataset. The anomalous attention map (in red) depicts the localization of the anomaly in the image. Please refer to the Sec. A.2 of Appendix for more illustrations of per class anomaly localization.

better, we observe from Table 3 that the mean of accuracy of correctly classified anomalous images and normal images of AVAGA$_u$ is better than the baseline methods. We achieve better classification performance in 10 out of 15 categories with an improvement ranging from 1% to 26% using AVAGA-D$_u$. In all baseline methods, the anomaly is localized from the thresholded pixel-wise difference between the input and reconstructed image where threshold is computed using anomalous images during training. It is important to note that in our unsupervised approach we do not use any anomalous images and generate an attention map to localize the anomaly which outperform the methods that have access to anomalous images. From Table 2, we observe that AVAGA-D$_w$ localizes the anomaly better than AVAGA-D$_u$ in all categories with an improvement ranging from 1% to 57%. We also observe that AVAGA-D$_w$ outperforms the baseline method with the best performance in 13 out of 15 categories with an improvement in localization between 1% and 45%. From Table 2 and Table 3, we observe that AE L2 and AE SSIM are the best performing methods for localization and classification accuracy as compared to other baseline approaches and hence choose to compare AVAGA$_u$ and AVAGA$_w$ with them on the modified STC dataset. We observe from Table 4 and Table 5 that AVAGA$_u$ and AVAGA$_w$ also outperforms these autoencoder based methods on modified STC dataset. Since we do not use any anomalous images in the unsupervised setting, we empirically set 0.5 as the threshold on attention map to evaluate the localization performance. Also, the anomalous score is normalized between $[0, 1]$, and 0.5 is empirically chosen as the threshold to detect an image as anomalous. From Table 7 in Appendix, we illustrate that AVAGA$_u$ is insensitive to the threshold and still outperforms the baselines methods for different threshold values.

Table 3: Comparison of mean of accuracy of correctly classified anomalous images and normal images of AVAGA$_u$ and AVAGA$_w$ with state-of-the-art approaches on the MvAD dataset. The representation of the highlighted number is the same as described in Table 2.

Category	AE SSIM	AE L2	AnoGAN	CNN feature dictionary	Texture inspection	Variation model	AVAGA-D$_u$	AVAGA-R$_u$	AVAGA-D$_w$	AVAGA-R$_w$
Bottle	0.88	0.80	0.69	0.53	-	0.57	0.89	0.91	0.93	0.96
Hazelnut	0.54	0.88	0.50	0.49	-	-	0.82	0.84	0.90	0.92
Capsule	0.61	0.62	0.58	0.41	-	0.50	0.81	0.87	0.89	0.93
Metal Nut	0.54	0.73	0.50	0.65	-	0.58	0.66	0.67	0.81	0.88
Leather	0.46	0.44	0.52	0.67	0.50	-	0.71	0.75	0.80	0.84
Pill	0.60	0.62	0.62	0.46	-	0.57	0.88	0.91	0.93	0.97
Wood	0.83	0.74	0.68	0.84	0.71	-	0.85	0.88	0.89	0.89
Carpet	0.67	0.50	0.49	0.63	0.59	-	0.71	0.78	0.80	0.82
Tile	0.52	0.77	0.51	0.71	0.72	-	0.70	0.72	0.81	0.86
Grid	0.69	0.78	0.51	0.67	0.50	-	0.75	0.78	0.79	0.81
Cable	0.61	0.56	0.53	0.61	-	-	0.62	0.64	0.86	0.97
Transistor	0.52	0.71	0.67	0.58	-	-	0.72	0.73	0.80	0.89
Toothbrush	0.74	0.98	0.57	0.57	-	0.80	0.90	0.97	0.96	0.99
Screw	0.51	0.69	0.35	0.43	-	0.55	0.77	0.78	0.79	0.79
Zipper	0.80	0.80	0.59	0.54	-	-	0.85	0.94	0.95	0.96

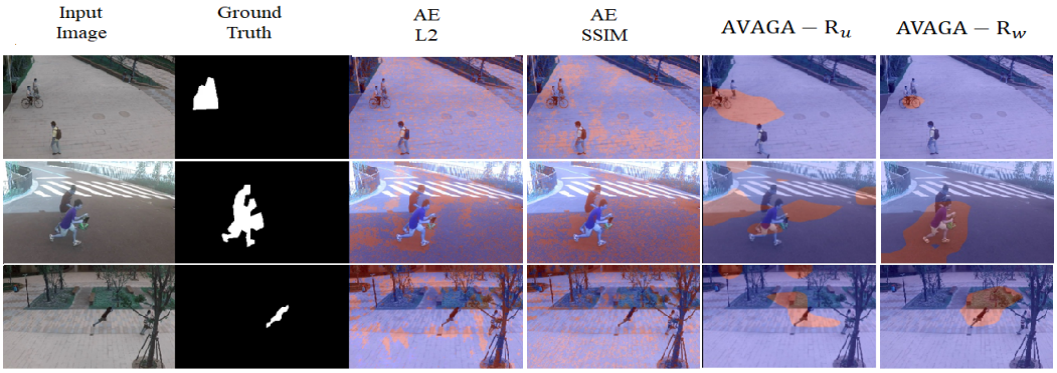

Figure 3: Qualitative results on the modified STC dataset. The anomalous attention map (in red) depicts the localization of the anomaly in the image. Please refer to the Sec. A.3 of Appendix for more illustrations of per scene anomaly localization.

Table 4: Comparison of IoU of AVAGA$_u$ and AVAGA$_w$ with state-of-the-art approaches on the modified STC dataset. The representation of the highlighted number is the same as described in Table 2.

Scene ID	AE SSIM	AE L2	AVAGA-D$_u$	AVAGA-R$_u$	AVAGA-D$_w$	AVAGA-R$_w$
01	0.20	0.16	0.20	0.25	0.38	0.44
02	0.08	0.17	0.19	0.23	0.25	0.34
03	0.21	0.24	0.26	0.28	0.31	0.46
04	0.11	0.12	0.28	0.34	0.36	0.38
05	0.16	0.12	0.28	0.31	0.40	0.47
06	0.21	0.19	0.31	0.40	0.45	0.58
07	0.19	0.16	0.18	0.22	0.28	0.36
08	0.06	0.05	0.20	0.22	0.29	0.37
09	0.03	0.02	0.20	0.16	0.31	0.36
10	0.11	0.14	0.13	0.14	0.24	0.29
11	0.10	0.07	0.29	0.37	0.44	0.58
12	0.20	0.16	0.07	0.11	0.20	0.26

5 Ablation Study

All ablation studies are performed on 5 randomly chosen categories of MvAD dataset. The quantitative and qualitative results are shown in Table 6 and Figure 4 respectively. The ablation results for all categories are presented in Table 8 of Appendix.

Effect of guided attention loss: To test the effectiveness of using the guided attention loss (L_{attn}) in the unsupervised setting, we train AVAGA-R$_u$ without it i.e. we use $L_R(x, \hat{x}) + KL(q_\phi(z|x)||p_\theta(z|x)) + L_{adv}$ as our objective function. During inference, the anomalous attention map is computed to evaluate the localization performance. From Column ID 1 & 3 in Table 6 , we observe that using guided attention loss localizes the anomaly better.

Effect of convolutional latent space: As discussed in Sec. 2.1.1, AVAGA$_u$ comprises of a convolutional latent space and to illustrate its effectiveness, we flatten the output of the encoder of AVAGA-R$_u$ and connect it to a fully connected layer as latent space with dimension 100. The dimension of the latent space is chosen from validation. From Column 2 & 3 in Table 6, we observe that preserving the spatial relation of the input and reconstructed image through the convolutional latent space results in a better localization performance.

Effect of adversarial reconstruction loss: We know that the attention map from the prediction of a trained classifier can be used to localize the object of interest corresponding to the classifier's prediction (Li et al. (2018)). We emphasize the effectiveness of training AVAGA-R$_w$ using eq. 6 as compared to training it only using a classification loss i.e. we train AVAGA-R$_w$ only using L_{bce} and then during inference, the anomalous attention map from the classifier's prediction localizes the anomaly on anomalous images. From Column ID 4 & 6 in Table 6 and the qualitative results illustrated in Figure 4, we observe that using the adversarial reconstruction loss jointly with guided attention loss results in a better localization performance.

Effect of SGB: From Sec. 2.2.1, we use SGB in AVAGA$_w$ to compute an attention map that localizes the anomaly better. We illustrate it's effectiveness by computing the attention loss for the image

Table 5: Comparison of mean of accuracy of correctly classified anomalous images and normal images of AVAGA$_u$ and AVAGA$_w$ with state-of-the-art approaches on the modified STC dataset. The representation of the highlighted number is the same as described in Table 2.

Scene ID	AE SSIM	AE L2	AVAGA-D$_u$	AVAGA-R$_u$	AVAGA-D$_w$	AVAGA-R$_w$
01	0.65	0.72	0.76	0.85	0.84	0.87
02	0.70	0.61	0.75	0.82	0.89	0.90
03	0.79	0.71	0.82	0.84	0.86	0.88
04	0.81	0.66	0.80	0.80	0.81	0.83
05	0.71	0.67	0.79	0.84	0.90	0.94
06	0.47	0.55	0.64	0.67	0.65	0.70
07	0.36	0.59	0.60	0.64	0.75	0.77
08	0.69	0.70	0.74	0.74	0.76	0.80
09	0.84	0.73	0.87	0.88	0.90	0.91
10	0.83	0.88	0.88	0.92	0.94	0.94
11	0.71	0.75	0.79	0.81	0.83	0.83
12	0.65	0.52	0.75	0.78	0.81	0.83

irrespective of the classifier's prediction. From Figure 4 and Column ID 5 & 6 in Table 6, we observe that using SGB to compute the guided attention loss results in better localization performance.

Table 6: IoU of 5 categories of ablation study illustrating the performance of the anomaly localization on MvAD dataset. Representation of highlighted number is same as described in Table 2.

Category	AVAGA-R$_u$ w/o attention	AVAGA-R$_u$ w/ flat latent space	AVAGA-R$_u$	AVAGA-R$_w$ w/o adv. recons	AVAGA-R$_w$ w/o SGB	AVAGA-R$_w$
Column ID	1	2	3	4	5	6
carpet	0.53	0.42	0.73	0.69	0.77	0.81
capsule	0.14	0.08	0.27	0.18	0.37	0.41
leather	0.18	0.31	0.79	0.72	0.81	0.84
Pill	0.16	0.25	0.38	0.24	0.44	0.53
Wood	0.43	0.36	0.59	0.51	0.61	0.66

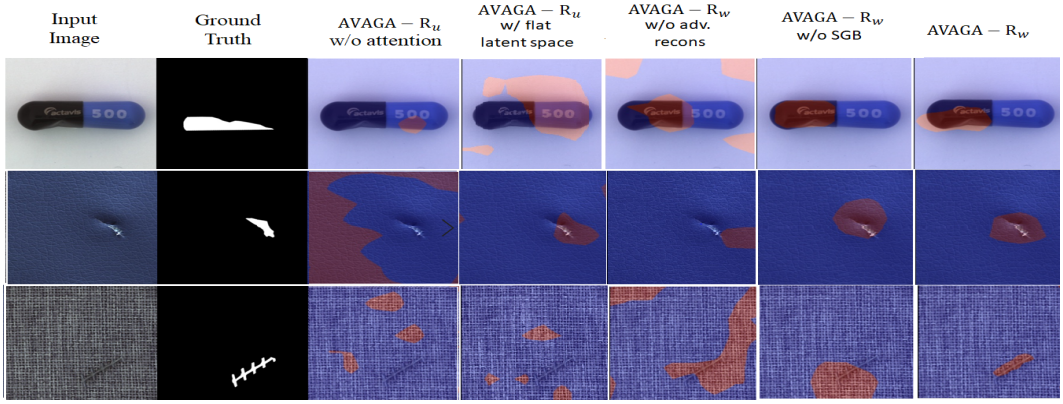

Figure 4: Qualitative results of the ablation study to illustrate the performance of the anomaly localization on MvAD dataset. The anomalous attention map (in red) depicts the localization of the anomaly for different cases as described in Sec. 5.

6 Conclusion

In this work, we propose the first end-to-end trainable convolutional adversarial variational autoencoder using guided attention to address anomaly detection and localization with attention maps. We illustrate that the guided attention loss during training enables the network to learn a feature representation of all the normal regions of the image such the anomalous attention map localize the anomaly. We also demonstrate that in the weakly supervised setting, using selective gradient backpropagation for guided attention along with 2% anomalous images during training improves the performance of anomaly localization. With qualitative and quantitative analysis the effectiveness of AVAGA$_u$ and AVAGA$_w$ is demonstrated over state-of-the-art methods for both unsupervised and weakly supervised settings on MvAD and modified STC datasets. Our proposed objective functions can be supported on different networks to improve the performance of detection and attention maps can be used for localization in both unsupervised and weakly supervised settings.

REFERENCES

Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian Conference on Computer Vision*, pp. 622–637. Springer, 2018.

Yannick Benezeth, P-M Jodoin, Venkatesh Saligrama, and Christophe Rosenberger. Abnormal events detection based on spatio-temporal co-occurences. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2458–2465. IEEE, 2009.

Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019.

Tobias Böttger and Markus Ulrich. Real-time texture error detection on textured surfaces with compressed sensing. *Pattern Recognition and Image Analysis*, 26(1):88–94, 2016.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Abnormal crowd behavior detection and localization using maximum sub-sequence search. In *Proceedings of the 4th ACM/IEEE international workshop on Analysis and retrieval of tracked events and motion in imagery stream*, pp. 49–58. ACM, 2013.

Asimenia Dimokranitou. *Adversarial autoencoders for anomalous event detection in images*. PhD thesis, 2017.

Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 244–253, 2019.

Matheus Gutoski, Nelson Marcelo Romero Aquino, Manassés Ribeiro, EA Lazzaretti, and SH Lopes. Detection of video anomalies using convolutional autoencoders and one-class support vector machines. In *XIII Brazilian Congress on Computational Intelligence, 2017*, 2017.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 733–742, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9215–9223, 2018.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545, 2018.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pp. 311–320. Springer, 2018.

Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Anomaly detection in nanofibrous materials by cnn-based self-similarity. *Sensors*, 18(1):209, 2018.

Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 685–694, 2015.

Nick Pawlowski, Matthew CH Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan Stevenson, Aneesh Khetani, Tom Newman, et al. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders. 2018.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Mahdyar Ravanbakhsh, Enver Sangineto, Moin Nabi, and Nicu Sebe. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1896–1904. IEEE, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pp. 146–157. Springer, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Carsten Steger. Similarity measures for occlusion, clutter, and illumination invariant object recognition. In *Joint Pattern Recognition Symposium*, pp. 148–154. Springer, 2001.

Du Tran and Junsong Yuan. Optimal spatio-temporal path discovery for video event detection. In *CVPR 2011*, pp. 3321–3328. IEEE, 2011.

Ha Son Vu, Daisuke Ueta, Kiyoshi Hashimoto, Kazuki Maeno, Sugiri Pranata, and Sheng Mei Shen. Anomaly detection with adversarial dual autoencoders. *arXiv preprint arXiv:1902.06924*, 2019.

Siqi Wang, En Zhu, Jianping Yin, and Fatih Porikli. Video anomaly detection and localization by local motion based joint video representation and ocelm. *Neurocomputing*, 277:161–175, 2018.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

A APPENDIX

A.1 DISCUSSIONS

Implementation details: All high resolution images of the MvAD and modified STC datasets are randomly center cropped to 256×256 and randomly rotated between $[-15°, +15°]$ to create variations in data during training. We train AVAGA$_u$ and AVAGA$_w$ with a learning rate of $1e^{-4}$ with a batch size of 24 for 150 epochs. In order to stabilize the training, the learning rate is decayed by $1e^{-1}$ for every 30 epochs. In the modified STC dataset, we train AVAGA$_u$ and AVAGA$_w$ with every 5^{th} frame of a video from each scene without using any temporal information.
Effect of varying threshold: The baseline methods shown in Table 2 use anomalous images to compute a threshold for detecting and localizing anomalies. However as discussed in Sec. 4, we empirically set 0.5 as our threshold for both detection and localization tasks. To illustrate that AVAGA-R$_u$ is insensitive to the variations in threshold values we choose 0.2 through 0.6 as our threshold on attention map. We compare the localization performance of AVAGA-R$_u$ for different threshold values to the best baseline method (SoTA best) in each category of MvAD dataset. From Table 7, we observe that AVAGA-R$_u$ is insensitive to the variations in threshold and still outperforms the best baseline method in each category.

Table 7: Comparision of IoU of AVAGA-R$_u$ for different threshold values with the best baseline method of each category (SoTA best) on MvAD dataset. The representation of the highlighted number is the same as described in Table 2.

Network Threshold	SoTA best	AVAGA-R$_u$ 0.2	AVAGA-R$_u$ 0.3	AVAGA-R$_u$ 0.4	AVAGA-R$_u$ 0.5	AVAGA-R$_u$ 0.6
Carpet	0.69	0.70	0.69	0.71	0.73	0.71
Capsule	0.11	0.23	0.24	0.22	0.27	0.18
Pill	0.25	0.34	0.31	0.35	0.38	0.33
Bottle	0.22	0.29	0.28	0.29	0.33	0.31
Wood	0.51	0.52	0.54	0.57	0.59	0.56
Tile	0.23	0.28	0.26	0.31	0.32	0.32
Hazelnut	0.41	0.41	0.45	0.46	0.47	0.46
Metal Nut	0.26	0.38	0.39	0.44	0.46	0.36
Cable	0.13	0.28	0.31	0.41	0.43	0.38
Toothbrush	0.51	0.52	0.52	0.53	0.55	0.53
Screw	0.34	0.38	0.44	0.45	0.48	0.46
Transistor	0.22	0.27	0.25	0.31	0.34	0.29
Zipper	0.13	0.19	0.20	0.24	0.25	0.22

Table 8: IoU of all categories of the ablation study illustrating the performance of the anomaly localization on MvAD dataset. The representation of the highlighted number is the same as described in Table 2.

Category	AVAGA-R$_u$ w/o attention	AVAGA-R$_u$ w/ flat latent space	AVAGA-R$_u$	AVAGA-R$_w$ w/o adv. recons	AVAGA-R$_w$ w/o SGB	AVAGA-R$_w$
Bottle	0.26	0.24	0.33	0.16	0.34	0.39
Hazelnut	0.16	0.26	0.47	0.51	0.76	0.79
Capsule	0.14	0.08	0.27	0.18	0.35	0.41
Metal Nut	0.28	0.31	0.45	0.25	0.38	0.46
Pill	0.16	0.25	0.38	0.24	0.44	0.53
Wood	0.43	0.36	0.59	0.51	0.61	0.66
Carpet	0.53	0.42	0.73	0.69	0.77	0.81
Tile	0.07	0.18	0.32	0.66	0.73	0.81
Leather	0.38	0.31	0.79	0.70	0.81	0.84
Grid	0.27	0.15	0.32	0.31	0.51	0.55
Cable	0.36	0.38	0.43	0.47	0.58	0.63
Toothbrush	0.41	0.46	0.55	0.54	0.60	0.66
Screw	0.11	0.18	0.48	0.16	0.22	0.31

Table 9: Architectural details of AVAGA$_u$ and AVAGA$_w$ as shown in Fig. 1. The notation in each row is as follows: operation, filter h $\times$ filter w, number of filters, stride, pad. W.S. denotes the additional layers for the weakly supervised setting. ConvTr 2D denotes transpose convolution layer, Conv 2D denotes convolution layer.

Network	Layer name	Layer dimensions	Output dimensions
Encoder	Layer 1 - 18	pretrained Resnet-18 (convolution only)	$8 \times 8 \times 512$
	Layer 19	ReLU	$8 \times 8 \times 512$
	Layer 20	Conv 2D, 1×1, 512, 1,0	$8 \times 8 \times 512$
	Layer 21	Conv 2D, 1×1, 512, 1,0	$8 \times 8 \times 512$
	W.S: Layer 22	Flatten	32768
	W.S: Layer 23	Linear	2
	W.S: Layer 24	Softmax	2
Decoder	Layer 1	ConvTr 2D, 4×4, 512, 2, 1	$16 \times 16 \times 512$
	Layer 2	BatchNorm	$16 \times 16 \times 512$
	Layer 3	ReLU	$16 \times 16 \times 512$
	Layer 4	Conv 2D 3×3, 512, 1, 1	$16 \times 16 \times 512$
	Layer 5	BatchNorm	$16 \times 16 \times 512$
	Layer 6	ReLU	$16 \times 16 \times 512$
	-	output layer 1 + output layer 6	$16 \times 16 \times 512$
	Layer 7	ConvTr 2D, 4×4, 256, 2, 1	$32 \times 32 \times 256$
	Layer 8	BatchNorm	$32 \times 32 \times 256$
	Layer 9	ReLU	$32 \times 32 \times 256$
	Layer 10	Conv 2D 3×3, 256, 1, 1	$32 \times 32 \times 256$
	Layer 11	BatchNorm	$32 \times 32 \times 256$
	Layer 12	ReLU	$32 \times 32 \times 256$
	-	output layer 7 + output layer 12	$32 \times 32 \times 256$
	Layer 13	ConvTr 2D, 4×4, 128, 2, 1	$64 \times 64 \times 128$
	Layer 14	BatchNorm	$64 \times 64 \times 128$
	Layer 15	ReLU	$64 \times 64 \times 128$
	Layer 16	Conv 2D 3×3, 128, 1, 1	$64 \times 64 \times 128$
	Layer 17	BatchNorm	$64 \times 64 \times 128$
	Layer 18	ReLU	$64 \times 64 \times 128$
	-	output layer 13 + output layer 18	$64 \times 64 \times 128$
	Layer 19	ConvTr 2D, 4×4, 64, 2, 1	$128 \times 128 \times 64$
	Layer 20	BatchNorm	$128 \times 128 \times 64$
	Layer 21	ReLU	$128 \times 128 \times 64$
	Layer 22	Conv 2D 3×3, 64, 1, 1	$128 \times 128 \times 64$
	Layer 23	BatchNorm	$128 \times 128 \times 64$
	Layer 24	ReLU	$128 \times 128 \times 64$
	-	output layer 19 + output layer 24	$128 \times 128 \times 64$
	Layer 25	ConvTr 2D, 4×4, 3, 2, 1	$256 \times 256 \times 3$
	Layer 26	Sigmoid	$256 \times 256 \times 3$
Discriminator	Layer 1	Conv2D, 4×4, 64, 2, 1	$128 \times 128 \times 64$
	Layer 2	Leaky ReLU (0.2)	$128 \times 128 \times 64$
	Layer 3	Conv2D, 4×4, 128, 2, 1	$64 \times 64 \times 128$
	Layer 4	BatchNorm	$64 \times 64 \times 128$
	Layer 5	Leaky ReLU (0.2)	$64 \times 64 \times 128$
	Layer 6	Conv2D, 4×4, 256, 2, 1	$32 \times 32 \times 256$
	Layer 7	BatchNorm	$32 \times 32 \times 256$
	Layer 8	Leaky ReLU (0.2)	$32 \times 32 \times 256$
	Layer 9	Conv2D, 4×4, 512, 2, 1	$16 \times 16 \times 512$
	Layer 10	BatchNorm	$16 \times 16 \times 512$
	Layer 11	Leaky ReLU (0.2)	$16 \times 16 \times 512$
	Layer 12	Conv2D, 4×4, 512, 2, 1	$8 \times 8 \times 512$
	Layer 13	Sigmoid	$8 \times 8 \times 512$

A.2 ADDITIONAL QUALITATIVE RESULTS - MVTEC ANOMALY DETECTION DATASET

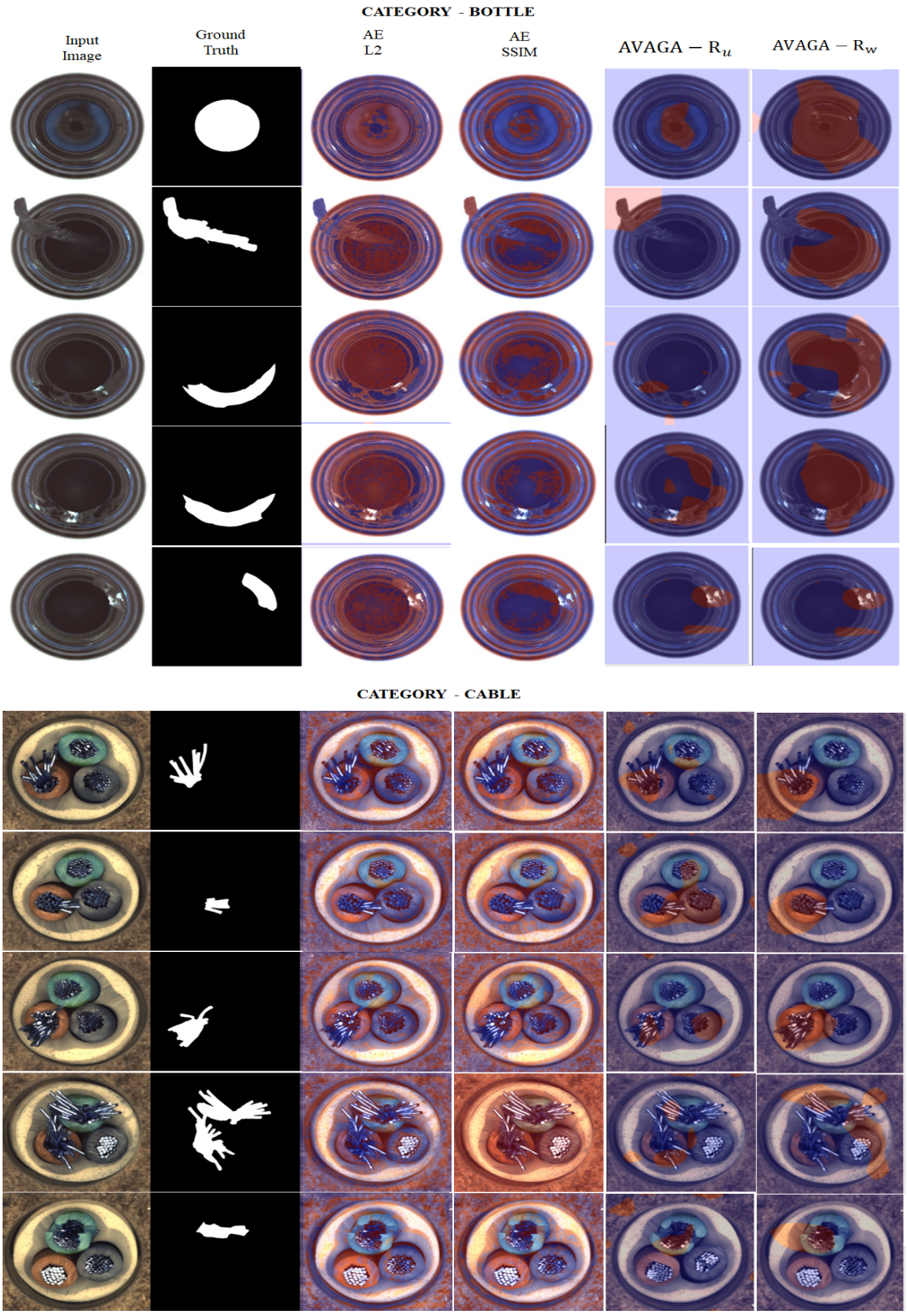

Figure 5: Qualitative comparison of anomaly localization of AVAGA-R_u and AVAGA-R_w with baseline methods on MvAD dataset. The anomalous attention map (in red) depicts the localization of the anomaly in the image.

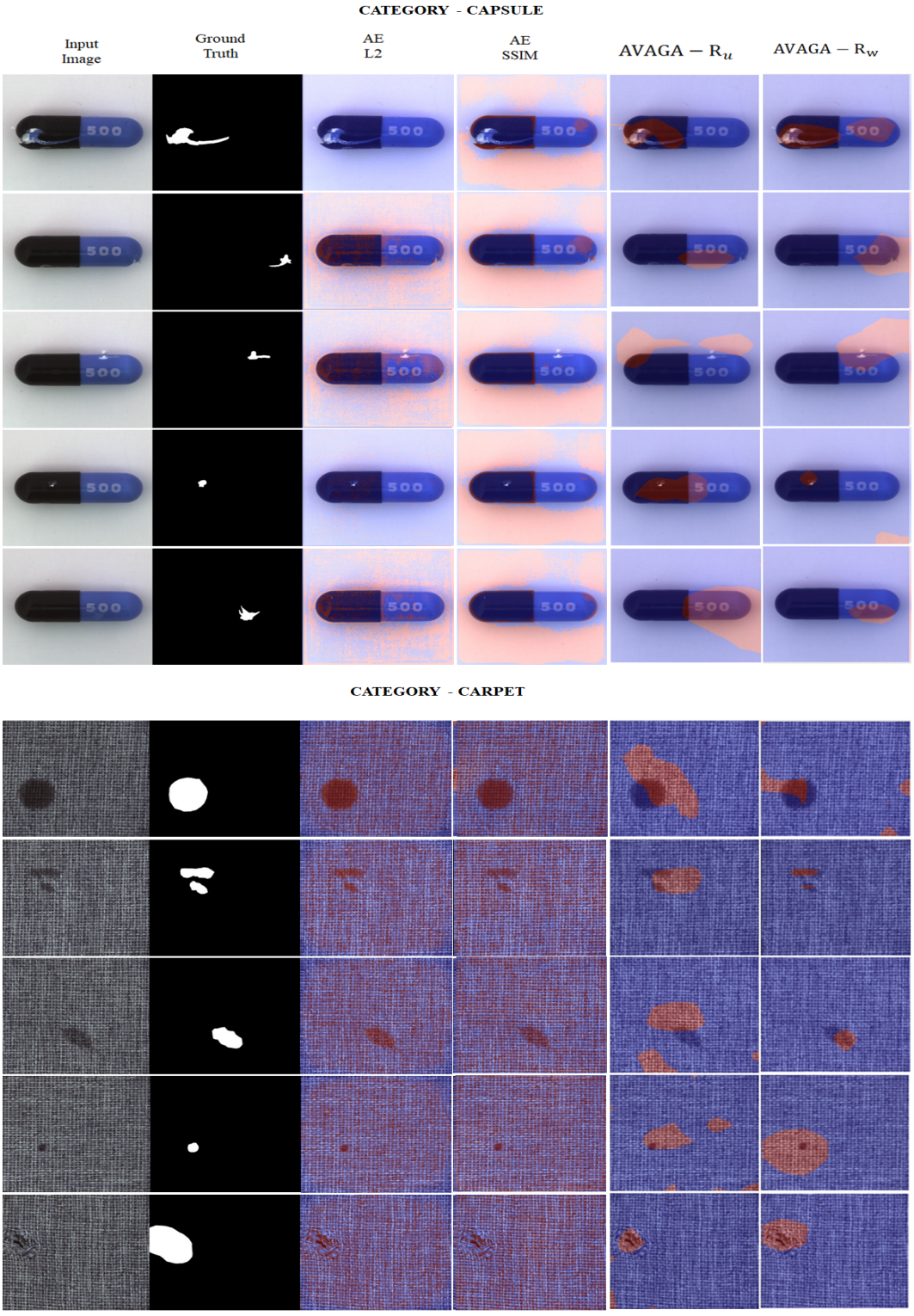

Figure 6: Qualitative comparison of anomaly localization of AVAGA-R_u and AVAGA-R_w with baseline methods on MvAD dataset. The anomalous attention map (in red) depicts the localization of the anomaly in the image.

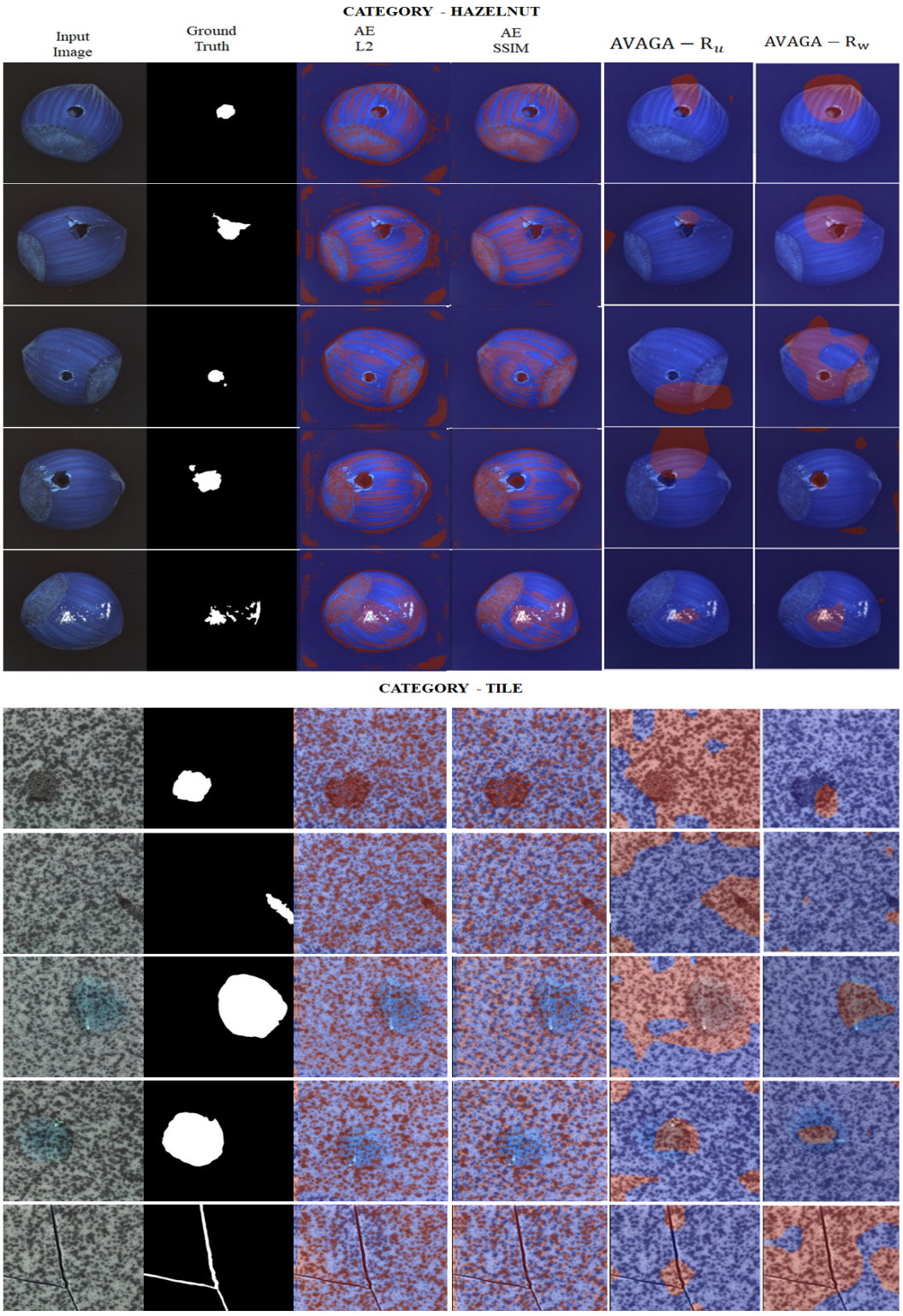

Figure 7: Qualitative comparison of anomaly localization of AVAGA-R_u and AVAGA-R_w with baseline methods on MvAD dataset. The anomalous attention map (in red) depicts the localization of the anomaly in the image.

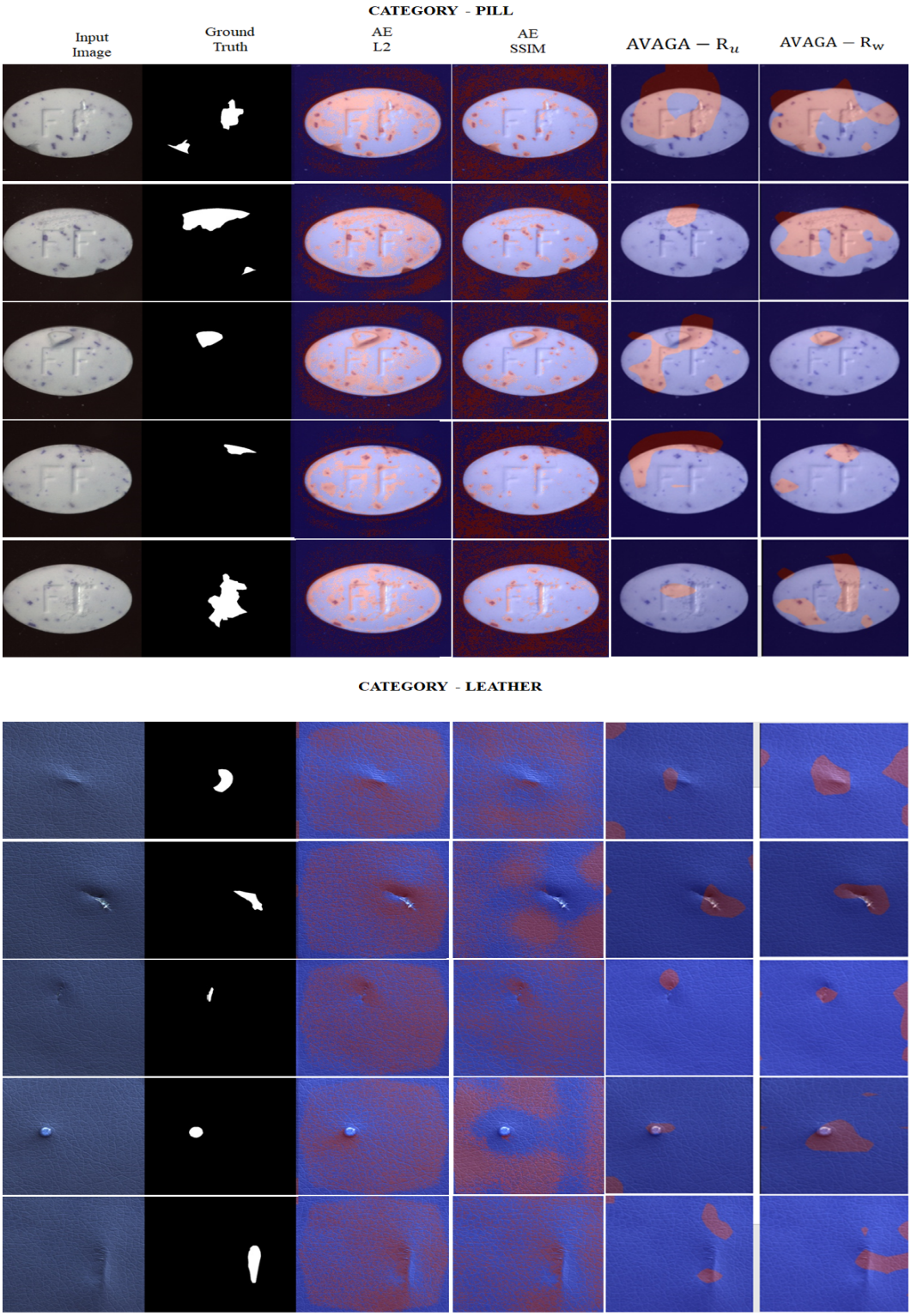

Figure 8: Qualitative comparison of anomaly localization of AVAGA-R_u and AVAGA-R_w with baseline methods on MvAD dataset. The anomalous attention map (in red) depicts the localization of the anomaly in the image.

A.3 ADDITIONAL QUALITATIVE RESULTS - SHANGHAITECH CAMPUS DATASET

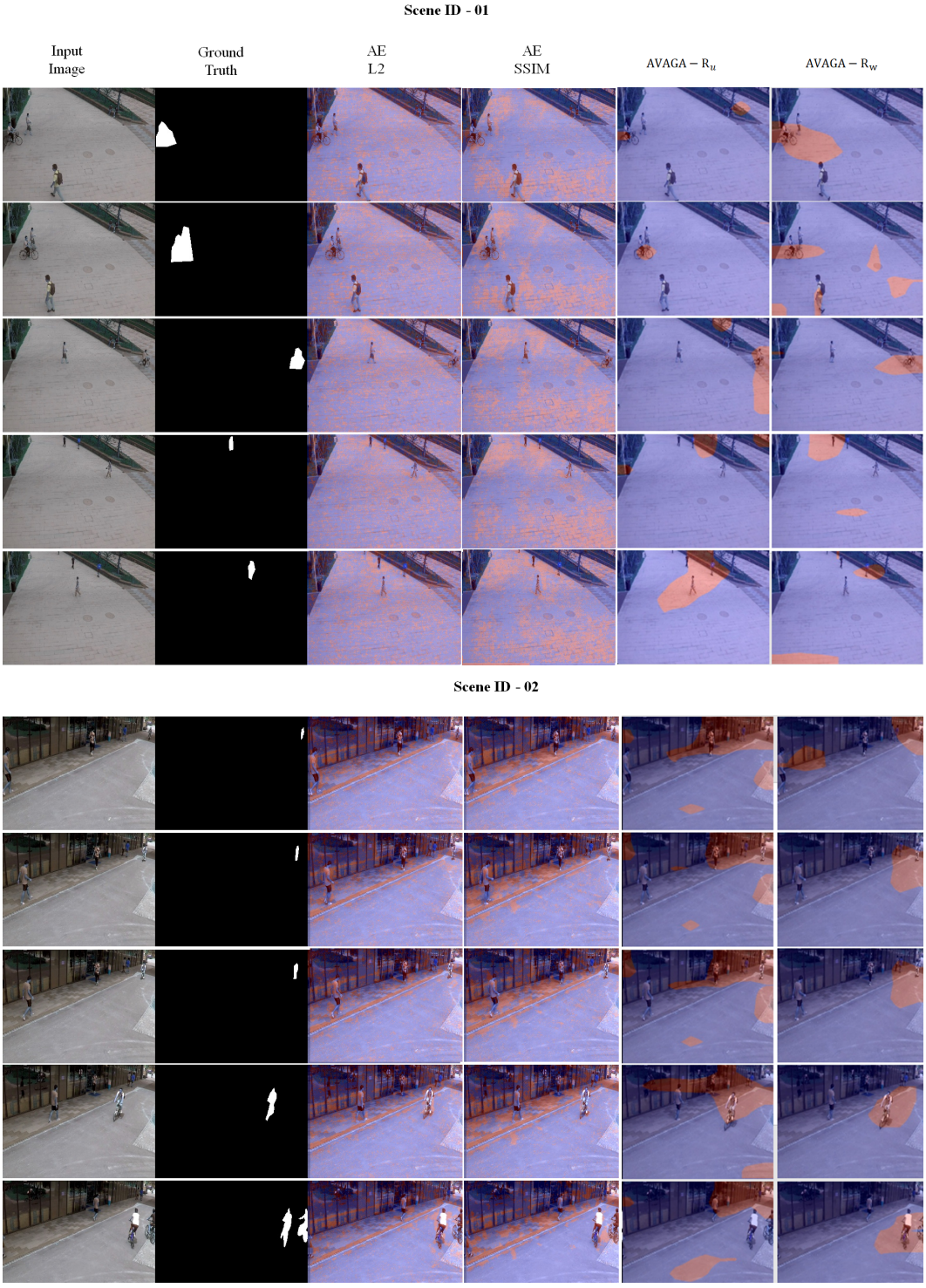

Figure 9: Qualitative comparison of anomaly localization of AVAGA-R_u and AVAGA-R_w with baseline methods on modified STC dataset. The anomalous attention map (in red) depicts the localization of the anomaly in the image.

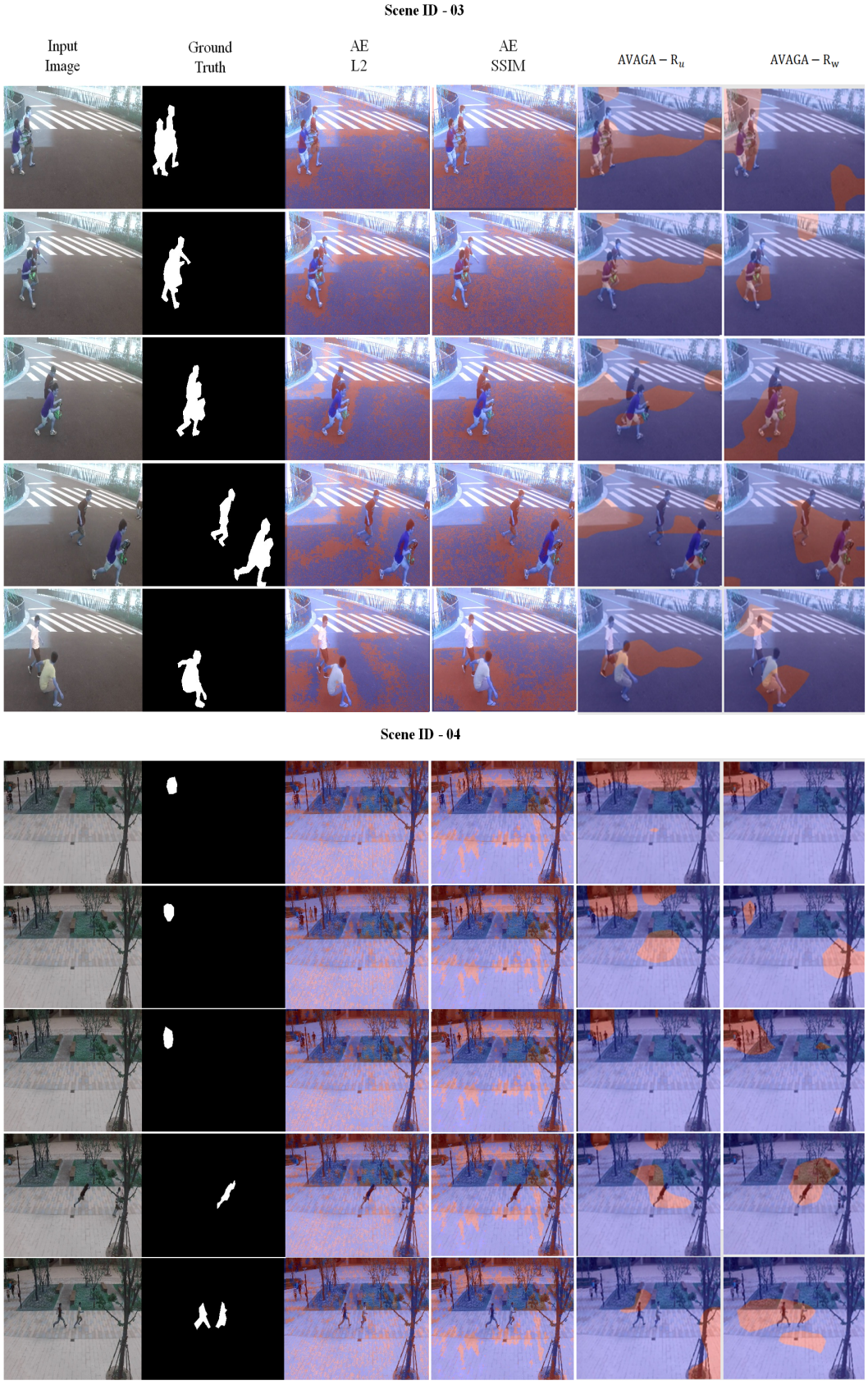

Figure 10: Qualitative comparison of anomaly localization of AVAGA-R_u and AVAGA-R_w with baseline methods on modified STC dataset. The anomalous attention map (in red) depicts the localization of the anomaly in the image.