Learning Unsupervised Metaformer for Anomaly Detection
Supplementary Material

Jihih-Ciang Wu[1,2], Ding-Jie Chen[1], Chiou-Shann Fuh[2], and Tyng-Luh Liu[1]
[1]Institute of Information Science, Academia Sinica, Taiwan
[2]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan

Network. The autoencoder $\mathcal{A}$ consists of an encoder and a decoder. The encoder is made up of convolutional layers. The decoder is symmetric to the encoder and composed of transposed convolutional layers. Each convolutional layer is followed by batch normalization, leaky ReLU activation, and a downsampling/upsampling layer. Details of the network are presented in Table 1.

Qualitative Analysis. The instance-prior generator is trained with MSRA10K and Flickr online images in an unsupervised manner. Figure 1 visualizes the response map derived from the generator $\mathcal{P}$, which can discriminate most parts of the foreground object for both MSRA10K and MVTec AD images. The response map R is then fed into the transformer $\mathcal{T}$ as a clue to predict anomalous regions. Figure 2 shows one example of all visual results in our metaformer. Figure 3 shows some qualitative examples, including two failure examples of the `cable` and `bottle` categories. The failures may be caused by the co-existing multiple anomalous types in a single image. For example, the cable instance in Figure 3 combines *missing cable* and *cable swap* defects, and the Metaformer only detects the former part. We believe that this situation will be mitigated once the gap between anomalous types in a single image is decreased like `wood` in Figure 3.

Table 1. Architecture for our autoencoder $\mathcal{A}$.

Layer	Output Size	Kernel
Input	$256 \times 256 \times 3$	-
Conv1	$256 \times 256 \times 32$	5×5
MaxPool	$128 \times 128 \times 32$	2×2
Conv2	$128 \times 128 \times 64$	5×5
MaxPool	$64 \times 64 \times 64$	2×2
Conv3	$64 \times 64 \times 128$	5×5
MaxPool	$32 \times 32 \times 128$	2×2
Conv4	$32 \times 32 \times 256$	5×5
MaxPool	$16 \times 16 \times 256$	2×2
Conv5	$16 \times 16 \times 512$	5×5
MaxPool	$8 \times 8 \times 512$	2×2
AvgPool	$2 \times 2 \times 512$	-
Upsample	$16 \times 16 \times 512$	8×8
T.Conv1	$16 \times 16 \times 256$	5×5
Upsample	$32 \times 32 \times 256$	2×2
T.Conv2	$32 \times 32 \times 128$	5×5
Upsample	$64 \times 64 \times 128$	2×2
T.Conv3	$64 \times 64 \times 64$	5×5
Upsample	$128 \times 128 \times 64$	2×2
T.Conv4	$128 \times 128 \times 32$	5×5
Upsample	$256 \times 256 \times 32$	2×2
T.Conv5	$256 \times 256 \times 3$	5×5

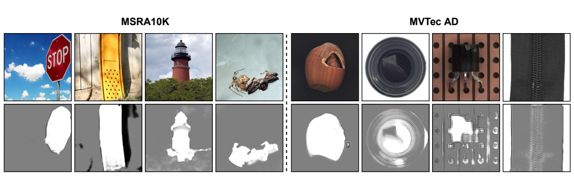

Figure 1. Visualization of the response map R obtains by instance-prior generator $\mathcal{P}$ for each dataset.

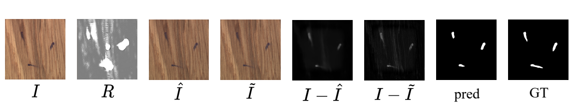

Figure 2. Example qualitative results.

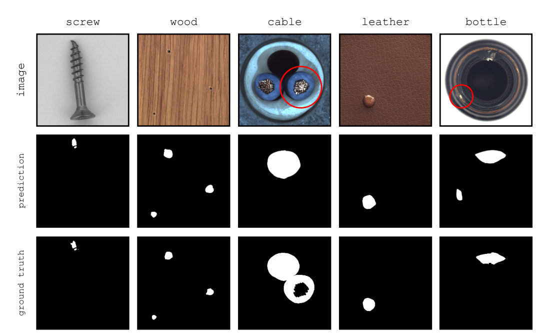

Figure 3. Qualitative analysis. We sample distinct categories from MVTec AD. The red circles indicate the miss-detect part.