Rethinking Reverse Distillation for Multi-Modal Anomaly Detection

Zhihao Gu†, Jiangning Zhang2, Liang Liu3, Xu Chen2, Jinlong Peng2, Zhenye Gan2, Guannan Jiang3, Annan Shu3, Yabiao Wang2, Lizhuang Ma1†

1School of Electronic and Electrical Engineering, Shanghai Jiao Tong University
2YouTu Lab, Tencent
3Contemporary Amperex Technology Co. Limited (CATL)

eIerry-holmes@sjtu.edu.cn, {vtzhang, cxxuchen, jeromepeng, wingzygan}@tencent.com, ma-lz@cs.sjtu.edu.cn

Abstract

In recent years, there has been significant progress in employing color images for anomaly detection in industrial scenarios, but it is insufficient for identifying anomalies that are invisible in RGB images alone. As a supplement, introducing extra modalities such as depth and surface normal maps can be helpful to detect these anomalies. To this end, we present a novel Multi-Modal Reverse Distillation (MMRD) paradigm that consists of a frozen multi-modal teacher encoder to generate distillation targets and a learnable student decoder targeting to restore multi-modal representations from the teacher. Specifically, the teacher extracts complementary visual features from different modalities via a siamese architecture and then parameter-freeley fuses these information from multiple levels as the targets of distillation. For the student, it learns modality-related priors from the teacher representations of normal training data and performs interaction between them to form multi-modal representations for target reconstruction. Extensive experiments show that our MMRD outperforms recent state-of-the-art methods on both anomaly detection and localization on MVTec-3D AD and Eyecandies benchmarks. Codes will be available upon acceptance.

Introduction

Anomaly detection (AD) has received continuous attention for several decades due to its wide range of applications such as defect detection, autonomous driving, video surveillance, and medical diagnosis. It is usually formulated as an unsupervised problem for the scarcity of anomalous data.

In recent years, vast efforts are dedicated to developing unsupervised anomaly detectors in images and tremendous progress has been made (Rudolph, Wandt, and Rosenhahn 2021; Roth et al. 2022; Li et al. 2021; Zavrtanik, Kristan, and Skočaj 2021; Hou et al. 2021; Deng and Li 2022), where embedding-based methods, synthesis and reconstruction are the dominant trends for this task. Embedding-based methods (Rudolph, Wandt, and Rosenhahn 2021; Roth et al. 2022) characterize the corresponding distribution of the extracted features, and the anomalies are detected by measuring the distance between features of test images and the estimated distribution. The synthesis-based methods (Li et al. 2021; Zavrtanik, Kristan, and Skočaj 2021) tend to utilize multiple modalities for AD, the necessity of which is illustrated in Fig. 2. The hole on the “cookie” and the protrusion on the “lollipop” is imperceptible on RGB images, but can be detected using depth and surface normals as the auxiliary modality. Besides, they also pro-

Figure 1: Illustration of different multi-modal anomaly detectors and corresponding anomaly maps (last row). Left: Reverse distillation. Middle: Two-stream structure with late fusion. Right: Our proposed paradigm.
Figure 2: First row: normal samples. Second row: defective samples. Depth and normals provide supplementary visual information to RGB images for revealing anomalies and reducing misidentification of anomaly-free areas.

The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)
In AD, the Knowledge Distillation (KD) detects anomalies based on RGB images and contains a pre-trained teacher network. It owns two types: 1) the Forward Distillation (FD) (Bergmann et al. 2020; Wang et al. 2021); 2) Reverse Distillation (RD) (Deng and Li 2022). Formally, given an RGB image \( I_R \in \mathbb{R}^{C \times H \times W} \) (\( C, H, \) and \( W \) is the channel, height, and width), the frozen teacher extracts feature \( \{ F_R^i \} _{i=1}^{K} \in \mathbb{R}^{C_i \times H_i \times W_i} \) (distillation targets) from its \( K \) stages and the student is trained to restore them, resulting in \( \{ F_S^i \} _{i=1}^{K} \in \mathbb{R}^{C_i \times H_i \times W_i} \). Differently, the student in FD encodes \( I_R \) but the student in RD decodes the one-class embedding of the teacher. Finally, a KD loss is used to supervise the reconstruction process:

\[
\mathcal{L}^{\text{KD}}_i = 1 - \frac{\text{flat}(F_S^i)^T}{\text{flat}(F_R^i)} \cdot \frac{\text{flat}(F_S^i)}{\|\text{flat}(F_S^i)\|_2},
\]

where \( \text{flat}() \) is the flatten function. In inference, pixel-wise cosine similarity between \( \{ F_R^i, F_S^i \} _{i=1}^{K} \) is computed to detect and localize anomalies, as shown in Fig. 1-Left.

However, it is difficult for KD to detect anomalies invisible in RGB images. To handle it, based on RD, we develop a novel multi-modal reverse distillation paradigm, which contains a frozen multi-modal teacher encoder and a learnable multi-modal student decoder.

Multi-Modal Teacher (MMT) Encoder

For the teacher encoder, we generate multi-modal distillation targets by integrating supplementary information from an auxiliary modality with an RGB image. As illustrated in Fig. 3-Left, we adopt a cross-statistics siamese teacher network to extract those information and a modality modulation module to parameter-free produce these targets.

Cross-Statistics Siamese Teacher Network. Fig. 2 shows that auxiliary modalities provide supplementary visual information to RGB images for revealing anomalies and reducing misidentification of anomaly-free areas. To model such supplementarity, we adopt a shared encoder, known as the siamese network, to extract features from the RGB image and the corresponding auxiliary modality, denoted as \( \{ F_i, F_A \} _{i=1}^{K} \). Nevertheless, the teacher network in KD is pre-trained on RGB images, and statistics stored in Batch Normalization layers (BNs) are shifted for the auxiliary.

and shift operations to exchange information between multi-modal features. MGAF (Kim, Jones, and Hager 2021) fuses motion features with that from detection via the cross-attention (Wang et al. 2018). In KD for multi-modal AD, we not only perform parameter-free modality modulation to form distillation targets in the teacher but also generate multi-modal representations to help the student better restore these targets.

Proposed Method

This section revisits knowledge distillation for anomaly detection as preliminaries. Then, details of the proposed frozen multi-modal teacher encoder and learnable multi-modal student decoder are presented one by one. The overall paradigm is shown in Fig. 3 and the algorithm table summarizing the proposed method is included in the supplementary material.

Preliminaries: Knowledge Distillation for AD

In AD, the Knowledge Distillation (KD) detects anomalies based on RGB images and contains a pre-trained teacher network and a learnable student network. It owns two types: 1) the Forward Distillation (FD) (Bergmann et al. 2020; Wang et al. 2021); 2) Reverse Distillation (RD) (Deng and Li 2022). Formally, given an RGB image \( I_R \in \mathbb{R}^{C \times H \times W} \) (\( C, H, \) and \( W \) is the channel, height, and width), the frozen teacher extracts feature \( \{ F_R^i \} _{i=1}^{K} \in \mathbb{R}^{C_i \times H_i \times W_i} \) (distillation targets) from its \( K \) stages and the student is trained to restore them, resulting in \( \{ F_S^i \} _{i=1}^{K} \in \mathbb{R}^{C_i \times H_i \times W_i} \). Differently, the student in FD encodes \( I_R \) but the student in RD decodes the one-class embedding of the teacher. Finally, a KD loss is used to supervise the reconstruction process:

\[
\mathcal{L}^{\text{KD}}_i = 1 - \frac{\text{flat}(F_S^i)^T}{\text{flat}(F_R^i)} \cdot \frac{\text{flat}(F_S^i)}{\|\text{flat}(F_S^i)\|_2},
\]

where \( \text{flat}() \) is the flatten function. In inference, pixel-wise cosine similarity between \( \{ F_R^i, F_S^i \} _{i=1}^{K} \) is computed to detect and localize anomalies, as shown in Fig. 1-Left.

However, it is difficult for KD to detect anomalies invisible in RGB images. To handle it, based on RD, we develop a novel multi-modal reverse distillation paradigm, which contains a frozen multi-modal teacher encoder and a learnable multi-modal student decoder.

Multi-Modal Teacher (MMT) Encoder

For the teacher encoder, we generate multi-modal distillation targets by integrating supplementary information from an auxiliary modality with an RGB image. As illustrated in Fig. 3-Left, we adopt a cross-statistics siamese teacher network to extract those information and a modality modulation module to parameter-free produce these targets.

Cross-Statistics Siamese Teacher Network. Fig. 2 shows that auxiliary modalities provide supplementary visual information to RGB images for revealing anomalies and reducing misidentification of anomaly-free areas. To model such supplementarity, we adopt a shared encoder, known as the siamese network, to extract features from the RGB image and the corresponding auxiliary modality, denoted as \( \{ F_i, F_A \} _{i=1}^{K} \). Nevertheless, the teacher network in KD is pre-trained on RGB images, and statistics stored in Batch Normalization layers (BNs) are shifted for the auxiliary.
modality. To mitigate this issue, we share the frozen convolutions for both modalities but maintain individual BNs for the auxiliary modality. Relevant statistics in these BNs are updated within several epochs with parameters of affine transformation unchanged, whose impacts are explored in Tab. 2 (b) and visualized in the supplementary material. In practice, we also adopt this strategy for RGB images. As a result, the extracted features are more modality-specific.

**Parameter-Free Modality Modulation.** Note that since the frozen teacher in KD provides deterministic distillation targets for a given input, the modality fusion should contain no learnable parameters. Besides, as discussed before, the auxiliary modality owns supplementary visual information to RGB images and is integrated for an auxiliary purpose. Therefore, not all information in $F^A_i$ is equally needed. To this end, we propose to estimate a fusion weight for $F^A_i$ to decide how much information is needed to be fused and then compensate $F^R_i$ with the selected information in a residual form. Concretely, we first exploit a normalization operation to generate the fusion weight $\alpha^A_i \in \mathbb{R}^{N \times H_i \times W_i}$:

$$\alpha(F^A_i) = \text{Sigmoid}(\frac{(F^A_i - \mu^A_i)^2}{(\sigma^A_i)^2 + 10^{-4}}),$$

where $\mu^A_i = \frac{1}{n_{w,c}} \sum F^A_i$ and $(\sigma^A_i)^2 = \frac{1}{n_{w,c}} \sum (F^A_i - \mu^A_i)^2$. Intuitively, the normalization operation helps reduce the disturbance from modality-specific information and better reflects the position-wise intensity. In practice, we find that $\alpha_i^A$ calculated from the sum of $F^R_i$ and $F^A_i$, denoted as $F_i^c$, performs better than $\alpha(F^A_i)$. It may be because $F_i^c$ contains more comprehensive information than individual ones and thus is a better indicator for the fusion weight. We give the visual effects in Fig. 4. Finally, the multi-modal teacher representation (distillation target) $F^T_i$ is formulated as:

$$F^T_i = F^R_i + \alpha(F_i) \cdot F^A_i.$$

$\alpha(F_i) \in [0, 1]$ flexibly controls the multi-modal information. Compared to $F^R_i$, the multi-modal $F^T_i$ pays more attention to objects and suppresses the effects from the background, which is investigated in the supplementary material.

**Analysis.** The devised siamese teacher encoder differs from AsymFusion (Wang et al. 2020b) in two aspects. First, ours extracts modality-specific features by a frozen architecture but their fully learnable structure instead encodes multi-modal features in each branch. Second, we parameter-freely fuse features of each modality to generate multi-modal distillation targets while they fuse features for further encoding.

**Multi-Modal Student (MMS) Decoder.** For the student, we incorporate multi-modal prior information to help restore distillation targets. To this end, we first generate priors for each modality via a modality-related priors generation module and then perform interaction on them to produce multi-modal priors via a multi-modal priors generation module, as shown in Fig. 3-Right.

**Modality-Related Priors Generation.** In KD, the student is expected to restore representations of the teacher encoder. Therefore, introducing information from the teacher to the student is helpful for the reconstruction. We then propose to learn a set of representative features (named “prototypes”) from the teacher representations of normal training data and generate modality-related priors to provide finer modal information. The prototypes are learned for both modalities and integrated via feature retrieval to generate priors for each modality. Formally, given the teacher representation of an RGB image $F^R_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ and $N$ prototypes $P^R = \{(P^R_j)_{j=1}^N\}$, the position-wise retrieval weight $W^R_i \in \mathbb{R}^{N \times H_i \times W_i}$ is measured as follows:

$$(W^R_i)_{j,h,w} = \frac{\exp(d((F^R_i)^{h,w}, (P^R_j)))}{\sum_{j=1}^{N} \exp(d((F^R_i)^{h,w}, (P^R_j)))},$$

where $(w, h)$ denotes spatial index and $d(\cdot, \cdot)$ is the cosine similarity. Aggregating $P^R_i$ with weights at each location of $W^R_i$ gives the reconstruction result $\hat{F^R}_i$:

$$(\hat{F^R}_i)^{w,h} = \sum_j (W^R_i)^{j,h,w} \cdot (P^R_j).$$

To ensure $P^R_i$ learns representative information, we propose to enforce the similarity between the teacher representation $F^T_i$ and the reconstruction $\hat{F^R}_i$ in the training phase:

$$\mathcal{L}^R = \frac{1}{HWC} \sum_{h,w,c} \| F^R_i - \hat{F^R}_i \|^2_2.$$

Note that Eq. (6) is applied to all normal training samples. Therefore, the learned $P^R_i$ contains normal information and is representative enough. This is why we call them “prototypes”. In inference, the teacher representation $F^R_i$ is used to generate the modality-specific priors $\hat{F^R}_i$ via Eq. (5).

For the auxiliary modality, we also learn a set of $N$ prototypes $P^A_i = \{(P^A_j^i)_{j=1}^N\}$ via a similar process, producing the loss $\mathcal{L}^A$ and priors $\hat{F^A}_i$.

**Multi-Modal Priors Generation.** Next, we aim to provide multi-modal prior information for the student to reconstruct the distillation target $F^T_i$. To achieve this, we perform multi-modality interaction between the modality-related $\tilde{F^R}_i$ and $\hat{F^A}_i$ to obtain a refiner representation. Since the auxiliary modality provides supplementary visual cues and the student is learnable, we use $\hat{F^A}_i$ to enhance $\tilde{F^R}_i$ through the intra and inter-modal interaction, as demonstrated in Fig. 3-Right. Specially, we first conduct the Channel Attention (CA) (Hu, Shen, and Sun 2018) on $\tilde{F^R}_i$ for intra-modal enhancement. Then the Spatial Attention (SA) map of size $\mathbb{R}^{1 \times H_i \times W_i}$ is generated from $\hat{F^A}_i$ via the MaxPooling $- \text{Conv}_{3 \times 3} - \text{Sigmoid}$ procedure. Finally, we perform inter-modal interaction by multiplying the enhanced $\tilde{F^R}_i$ with the SA map to highlight locations of interest, resulting in a finer multi-modal representation $\hat{F^R}_i$. The whole multi-modal interaction process can be formulated as follows:

$$\hat{F^R}_i = \text{SA}(\hat{F^A}_i) \cdot (\text{CA}(\hat{F^R}_i) \cdot \tilde{F^R}_i + \hat{F^R}_i) + \hat{F^R}_i.$$
Finally, \( \hat{F}_I^R \) is concatenated with \( F_{I-1}^S \) as the input of the student decoder \( D_{I-1} \) to restore \( F_{I-1}^T \), resulting in \( F_{I-1}^S \):

\[
F_{I-1}^S = D_{I-1}([F_{I-1}^S; \hat{F}_I^R]).
\] (8)

The \( F_{I-1}^S \) and \( F_{I-1}^T \) are used to compute the distortion loss in Eq. (1) during training and detect anomalies in inference.

Analysis. We give some theoretical explanations on scores from priors. The student is trained to produce anomaly-free features and then anomaly-free areas are inside the convex combination of “prototypes”. Finally, anomalies fail to be inside the combination and features between the teacher and student have a higher reconstruction error. This insight is used for anomaly localization. Fig. 5 verifies the analysis.

Loss Function and Anomaly Detection

Loss Function. It consists of the distortion loss from \( K \) stages and the prototype learning loss of each modality:

\[
L = \sum_{i=1}^{K} L_{i}^{K,D} + \lambda \sum_{i=1}^{K} (L_{i}^{R} + L_{i}^{A}),
\] (9)

where \( K = 3 \) and \( \lambda \) is the balance factor, set 0.1 by default.

Anomaly Detection. In inference, pixel-wise cosine similarity between \( \{F_{I}^{T}, F_{I}^{S}\}_{i=1}^{K} \) is computed and then a bilinear up-sampling operation \( U_p(\cdot) \) is conducted to generate an anomaly map \( S_i \). The final anomaly map \( A \) is given by:

\[
A = g(\sum_i U_p(1 - d(F_{i}^{T}, F_{i}^{S}))),
\] (10)

where \( g(\cdot) \) denotes the Gaussian filter (Roth et al. 2002). \( A \) gives the localization results and a larger score on it indicates a higher probability of anomaly. We simply take its maximum value as the image-level anomaly score.

Experiments

Experimental Settings

Datasets. We conduct experiments on two multi-modal benchmarks, i.e., the MTVec 3D-AD (Bergmann et al. 2022) and the Eyecandies (Bonfiglioli et al. 2022). The former contains 4,147 scans captured from 10 object categories and provides modality of RGB images and Point Clouds (PCs). The latter consists of 10 categories with 1,500 samples for each type and provides RGB images, depth maps, and surface normals. Pixel-level annotations are available in both datasets to evaluate the anomaly localization performance.

8449
The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Baseline Methods. We compare ours with several SOTA multi-modal detectors, i.e., AST (Rudolph et al. 2023) using depths and RGBs, M3DM (Wang et al. 2023) using PCs and RGBs, PatchCore (Roth et al. 2022) with FPFH (Rusu, Blodow, and Beetz 2009) using PCs and RGBs, and Eye- candy (Bonfiglioli et al. 2022) using normals and RGBs.

Evaluation Metrics. The Area Under the Receiver Operator Curve (AUROC) and Precision Recall (AUPR) are used to quantify anomaly detection and localization capacity. The Per-Region Overlap (PRO) is also adopted for localization.

Implemental Details. Images are resized into 256 × 256 and Adam is used as the optimizer with a learning rate of 0.001. The model is trained for 400 epochs of batch size 16. the number of prototypes is set 50. The teacher network is a pre-trained WideResNet50 and the student is the same as RD. We adopt the depth and normals as auxiliary modalities for MVTec 3D-AD and Eyecandies datasets, respectively.

Main Results

Results on the MVTec 3D-AD. Tab. 1 (a) shows experimental results for anomaly detection using 3D data, RGB images, or their combination on the MVTec 3D-AD dataset. Image-level AUROC and pixel-level PRO for all classes are reported. First, we find that by solely relying on RGB images for detection, our method outperforms all 3D-based counterparts (with improvements of 2.4% on AUROCAD and 3.8% on PROAL) in terms of mean values. This is likely due to the complexity of 3D data and the limited efforts put into its development. However, it is also observed that geometric information in some targets, e.g., foam and peach, play a more important role in detecting anomalies (86.5% versus 84.7% on foam, and 94.7% versus 91.3% on peach) since these anomalies are visually underperceived in the 2D view. Finally, integrating 3D information gives larger improvements.

Results on the Eyecandies. The proposed method is also evaluated on the Eyecandies dataset and image-level AUROC and pixel-level PRO for all classes are reported in Tab. 1 (b). We observe that the overall performance on normals is higher than that on RGB images. This is because the normals describe the geometric shape of the target object and some geometric anomalies that are hard to be perceived from images can thus become visually identifiable, as demonstrated in Fig. 2. Additionally, introducing the normals to images further improves the performance. Compared to methods such as AST and Eyecandy that fuse multiple modalities via concatenation, our strategy performs feature-level fusion, surpassing them by a clear margin.

More comprehensive results. In Tab. 3, our method outperforms AST and M3DM in four out of five AD metrics. Moreover, it consumes less training time and the inference speed is 1-4x compared to AST and 10x compared to M3DM, demonstrating both the effectiveness and efficiency.

Ablation Study

Study on key components. We study the effectiveness of the multi-modal teacher (MMT) and two key components to methods such as AST and Eyecandy that fuse multiple modalities via concatenation. Instead, introducing an auxiliary modality to the teacher brings a large improvement (7.5% † on AUROCAD and 2.6% † on PROAL). For the student, generating modality-related priors from normal samples and con-

Table 2: Ablation study on the Eyecandies dataset. “PG”, “MMFD”, “SEM”, “CSA” and “SSA” refer to the modality-related prior generation, multi-modal forward distillation, SE module, channel, and spatial self-attention, respectively.

<table>
<thead>
<tr>
<th>Component</th>
<th>ROCAD</th>
<th>ROCAL</th>
<th>PRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>84.0</td>
<td>96.7</td>
<td>88.9</td>
</tr>
<tr>
<td>+MMT</td>
<td>91.5</td>
<td>97.2</td>
<td>91.5</td>
</tr>
<tr>
<td>+MMT+PG</td>
<td>92.6</td>
<td>97.6</td>
<td>92.0</td>
</tr>
<tr>
<td>All</td>
<td>94.0</td>
<td>98.3</td>
<td>93.6</td>
</tr>
</tbody>
</table>

(a) Study on key components.

<table>
<thead>
<tr>
<th>Modality</th>
<th>ROCAD</th>
<th>ROCAL</th>
<th>PRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>74.5</td>
<td>90.8</td>
<td>86.2</td>
</tr>
<tr>
<td>Normals</td>
<td>89.5</td>
<td>96.0</td>
<td>90.6</td>
</tr>
<tr>
<td>RGB</td>
<td>86.5</td>
<td>94.5</td>
<td>89.4</td>
</tr>
<tr>
<td>+Depth</td>
<td>92.8</td>
<td>97.2</td>
<td>91.3</td>
</tr>
<tr>
<td>+Normal</td>
<td>94.0</td>
<td>98.3</td>
<td>93.6</td>
</tr>
<tr>
<td>All</td>
<td>94.4</td>
<td>98.8</td>
<td>93.9</td>
</tr>
</tbody>
</table>

(b) Study on individual BNs.

<table>
<thead>
<tr>
<th>Method</th>
<th>ROCAD</th>
<th>ROCAL</th>
<th>PRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSA</td>
<td>91.2</td>
<td>97.9</td>
<td>92.4</td>
</tr>
<tr>
<td>SSA</td>
<td>92.5</td>
<td>98.1</td>
<td>91.6</td>
</tr>
<tr>
<td>α = 1</td>
<td>93.0</td>
<td>98.1</td>
<td>92.8</td>
</tr>
<tr>
<td>α(FAi)</td>
<td>93.4</td>
<td>98.2</td>
<td>93.1</td>
</tr>
<tr>
<td>α(Fj)</td>
<td>94.0</td>
<td>98.3</td>
<td>93.6</td>
</tr>
<tr>
<td>SEM(Fj)</td>
<td>90.2</td>
<td>97.0</td>
<td>92.6</td>
</tr>
</tbody>
</table>

(c) Study on distillation paradigms.

<table>
<thead>
<tr>
<th>Method</th>
<th>ROCAD</th>
<th>ROCAL</th>
<th>PRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>FD</td>
<td>70.8</td>
<td>85.6</td>
<td>78.0</td>
</tr>
<tr>
<td>MMFD</td>
<td>82.5</td>
<td>90.2</td>
<td>84.4</td>
</tr>
<tr>
<td>RD</td>
<td>84.0</td>
<td>96.7</td>
<td>88.9</td>
</tr>
<tr>
<td>MMRD</td>
<td>94.0</td>
<td>98.3</td>
<td>93.6</td>
</tr>
</tbody>
</table>

(d) Study on modality-related priors.

<table>
<thead>
<tr>
<th>Method</th>
<th>ROCAD</th>
<th>ROCAL</th>
<th>PRAD</th>
<th>PRAL</th>
<th>GPUH/FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AST†</td>
<td>93.7</td>
<td>97.5</td>
<td>97.4</td>
<td>33.7</td>
<td>94.6</td>
</tr>
<tr>
<td>M3DM†</td>
<td>93.6</td>
<td>99.2</td>
<td>97.7</td>
<td>43.9</td>
<td>96.2</td>
</tr>
<tr>
<td>Ours</td>
<td>95.0</td>
<td>99.2</td>
<td>98.1</td>
<td>42.1</td>
<td>97.6</td>
</tr>
</tbody>
</table>

(e) Study on fusion strategies.

(f) Study on number of prototype.

Table 3: More comprehensive results on the MVTec 3D-AD dataset. AD and AL are short for anomaly detection and localization. † means re-implementation. “GPUH” and FPS refer to GPU hours and frame per second, respectively.
ducting multi-modal interaction give improvements of different degrees. Finally, combining them all performs best.

**Study on individual BNs in MMT.** They are used to learn modality-related statistics for adaption and their impacts are listed in Tab. 2 (b). Adopting individual BNs benefits both anomaly detection and localization while applying them to surface normals alone contributes less to final results than to RGB images, implying that the network may have difficulty further adapting Image-Net pre-trained convolutions to other modalities. Visualizations in the supplementary material show that learning RGB-related information helps the pre-trained convolutions better describe anomalies, resulting in finer multi-modal representations for the teacher.

**Study on distillation paradigms.** We explore the generalization of our multi-modal strategies to the Forward Distillation (FD) and Reverse Distillation (RD), as listed in Tab. 2 (c). How to apply them to FD can be found in the supplementary material. It is observed that integrating an auxiliary modality to the RGB data via our strategies gives consistent improvement to different distillation paradigms, which implies the flexibility and expandability of our method.

**Study on different modalities.** Tab. 2 (d) studies the effects of different modalities and how to extend our method to more modalities can be found in the supplementary material. First, compared to depth, both normals and images provide useful information for AD and thus achieve better results. Second, fusing RGB data with depth or normals all bring significant improvement whereas the normals own larger gains (6.3% v.s. 7.5% on AUROC_{AD}, 2.7% v.s. 3.8% on AUROC_{AL} and 1.9% v.s. 4.2% on PRO_{AL}). Instead, integrating depth into images and normals produces limited improvement since depth introduces minor extra information.

**Study on different fusion strategies for F_i^T.** In Tab. 2 (e), we explore different ways to generate the multi-modal representation F_i^T, including the parameter-free Channel Self-Attention (CSA) and Spatial Self-Attention (SSA) (Wang et al. 2018), and the learnable SE Module (Hu, Shen, and Sun 2018) (SEM). We observe that no learnable transformations in CSA and SSA result in inaccurate attention computation and thus lead to unsatisfactory results. Besides, they can only handle two modalities. Surprisingly, element-wise addition between F_i^R and F_i^A (α = 1) outperforms above strategies. Contrarily, fusion with adaptive weight α produces better results, indicating that not all the information in the auxiliary modality is important. The SEM instead underperforms the vanilla addition. We guess the parameterized SEM produces unstable representations for the teacher.

**Study on number of prototypes.** The number of prototypes controls the amount of normal information to be learned for each modality, which is explored in Tab. 2 (f). We find that learning normal information benefits both anomaly detection and localization. And more prototypes lead to better detection while owning similar AUROC_{AL}. Instead, a larger N_i leads to more parameters and optimization difficulty, resulting in more performance drops. For the sake of higher localization results, we adopt N_i = 50 by default.

**Visualization Analysis**

**Sources for generating fusion weight α.** Note that α in Eq. (3) can also be obtained from F^A. To explore the difference, Fig. 4 visualizes α on the depth map and their corresponding anomaly maps on the image. As shown in Fig. 4 (b) and (d), α(F) highlights not only anomalous regions which are visible in auxiliary modality but also some regions with special patterns in RGB (the chocolate on the “cookie”). In this sense, α(F^A) fails to introduce auxiliary modality information in special pattern regions and leads to wrong results in Fig. 4 (c). On the contrary, α(F) enables the model to consult the composite information in special pattern regions and get a more accurate anomaly map in Fig. 4 (e).

**How multi-modal priors work?** To investigate it, we visualize its impacts in Fig. 5. The multi-modal priors suppress responses to normal patterns in both anomaly-free and anomalous samples, e.g., the chocolate on the “cookie” and the hollow on the “potato”. This is mainly because the multi-modal priors contain normal information and are trained to help the student decoder restore anomaly-free features. Therefore, anomalous regions are highlighted and responses to normal patterns are mitigated after calculating pixel-wise feature similarity between the teacher and student networks.

**Conclusion**

We present a novel MMRD paradigm for anomaly detection, which integrates an auxiliary modality into RGB images for better detection. It uses a frozen multi-modal teacher encoder to generate multi-modal distillation targets for the learnable student decoder to restore. As a result, it achieves superior results on two multi-modal benchmarks.
Acknowledgments

This research is supported in part by National Natural Science Foundation of China (No. 61972157 and 72192821), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200), Shanghai Sailing Program (22YF1420300 and 23YF1410500), CCF-Tencent Open Research Fund (RAGR20220121) and Young Elite Scientists Sponsorship Program by CAST (2022QNRC001).

References


Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; and Wandt, B. 2023. Asymmetric Student-Teacher Networks for Industrial


