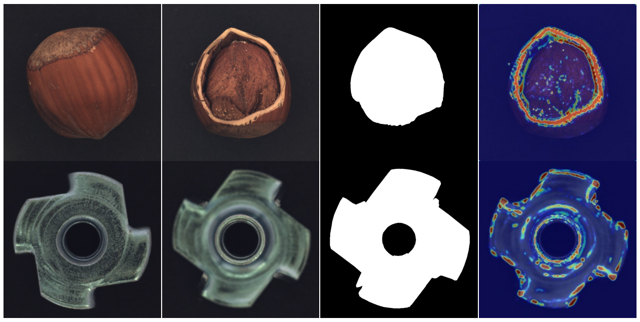

(a) Normal (b) Anomalous (c) Label (d) Prediction

Figure 3: Ambiguity of label. The diffusion model focuses on the anomalous pixels that need to be altered for successful reconstruction. It requires semantic information to correctly segment large-area hazelnut cracking and metal nut flipping.

in anomaly recall. As demonstrated in Fig. 3, the anomaly regions with a similar color to the normal pixels are assigned with a high likelihood by the denoising model. The diffusion model prioritizes the anomalous pixels that need to be altered for successful reconstruction, which requires semantic information to address the issue.

To enhance the accuracy of anomaly detection, we propose a joint distribution approach that considers both the pixel space and feature space, represented by $P(\boldsymbol{x}, \boldsymbol{f})$. We employ a pre-trained feature extractor to extract the deep features of the input image. The diffusion model is trained to concurrently reconstruct the pixels and semantic features of the noise-free image with the corrupted image. We adopt the Mean Squared Error (MSE) loss function as the training loss and anomaly score, which is defined as follows:

$$s_t^f = L_{mse}^f = \frac{1}{C \times H \times W} \sum |f(\boldsymbol{x}_0) - f(\boldsymbol{x}_t)|^2, \quad (10)$$

where f is a pre-trained feature extractor to extract features with shape $\mathbb{R}^{C \times H \times W}$, $\boldsymbol{x}_0$ and $\boldsymbol{x}_t$ represent a noise-free image and the corresponding corrupted image with random noises, respectively. The final anomaly score is the weighted sum of the pixel-level and feature-level results.

Multi-scale noises. We have observed that different anomalies exhibit varying sensitivities to different noise scales. While some anomalies can be detected easily, others require sufficiently large noise to overwhelm the anomalous pixels. We measure the anomaly score for various noise scales and average the results. Since the KL-divergence score varies significantly with the timestep t, we normalize it before averaging. The final anomaly score is obtained as follows:

$$A = \sum_{i=1,2,\ldots,n} \alpha \hat{s}_{t_i} + (1-\alpha) s_{t_i}^f, \quad (11)$$

Algorithm 1: Gradient Denoising Reconstruction.

```
Input: Image $\boldsymbol{x}_0$, Gaussian $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Output: $\boldsymbol{x}_N$
for $t = 1, \cdots, N$ do
    $\boldsymbol{f}_t = F(\boldsymbol{x}_t)$
    $\boldsymbol{g} = \nabla_{\boldsymbol{x}_t} (\boldsymbol{f}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{f}_t - \boldsymbol{\mu})$
    $\boldsymbol{x}_t = \sqrt{1 - \hat{\beta}_t \boldsymbol{x}_{t-1}} + \sqrt{\hat{\beta}_t \boldsymbol{g}}$ ;
    if $t \% N_d = 0$ then
        $\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t), \boldsymbol{\sigma}_\theta(\boldsymbol{x}_t))$
    end
end
```

where $\hat{s}_{t_i}$ is the normalized score by mean and standard deviation, $T = \{t_1, t_2, \cdots, t_n\}$ are the selected timesteps of the forward-process of the diffusion model. We analyze the effects of ensembling factor α in Sec. 4.4.

Unified model. It has been proved that the diffusion model's capacity of the diffusion network is large enough for modeling any complex distributions [15, 10]. Like UniAD [36], we conduct experiments to learn distributions of multiple categories with a single diffusion model. Table 3 illustrates that the performance of our unified model outperforms the other methods by a large margin under the single unified model setting. The results confirm the effectiveness of utilizing the diffusion model for anomaly localization.

3.3. Gradient Denosing for Reconstruction

An image's anomalous regions can be viewed as a special type of noise that can be removed using the diffusion model. We propose a gradient denoising process to remove the anomalies with simple adjustments to the reverse diffusion process of DDPM [15]. An anomalous image can be smoothly transformed into a normal one, providing an interpretable explanation of the anomaly detection results.

We first introduce a gradient descending optimization process for anomaly reconstruction. We take the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ approximated by PaDiM [8] on the deep features of anomaly-free data. For reconstruction, we extract embedding $f(\boldsymbol{x}_0)$ with the feature extractor of PaDiM and use the Mahalanobis distance to optimize the image with gradient descending:

$$L = (f(\boldsymbol{x}_0) - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (f(\boldsymbol{x}_0) - \boldsymbol{\mu}), \quad (12)$$

$$\boldsymbol{x}_{t+1} = \omega \boldsymbol{x}_t - s \nabla_{\boldsymbol{x}_t} L, \quad (13)$$

where ω is weight decay factor and s is the learning rate. The process optimizes the image such that the anomaly score of PaDiM is minimized. However, the noisy gradients $\nabla_{\boldsymbol{x}_t} L$ will corrupt the image after some iterations, introducing significant noises to the image. We propose to leverage the diffusion model to denoise the gradients for high-quality reconstruction.