

# Debiasing Trace Guidance: Top-down Trace Distillation and Bottom-up Velocity Alignment for Unsupervised Anomaly Detection

Xingjian Wang

Zhejiang University

xingjianwang@zju.edu.cn

Li Chai \*

Zhejiang University

chaili@zju.edu.cn

Jiming Chen

Zhejiang University

cjm@zju.edu.cn

## Abstract

*The leak of anomalous information from input condition poses a great challenge to reconstruction-based anomaly detection. Recent diffusion-based methods respond to this issue by suppressing anomaly information for condition injection or in-sampling inversion. However, since they treat conditions as a time-invariant prior, they fall into a trade-off problem between anomaly suppression and normal pattern consistency. To address this problem, we propose Debiasing Trace Guidance (DTG) framework based on Flow Matching towards debiasing generation for more accurate unsupervised multi-class anomaly detection. Generally, DTG distills a low-dimensional generation sub-trace robust to anomalies by Top-down Trace Distillation, and then utilizes its time-varying velocity features to guide a debiasing generation by Bottom-up Velocity Alignment. The trace distillation filters out high-frequency anomalies via learnable wavelet filters and reserving structural information by keeping global consistency across samples using Skin-horn Distance. Subsequently, the velocity field of original trace is aligned with the one of sub-trace through KV-Injection Attention mechanism. The model is forced to generate normal details from corresponding low-dimensional contexts via Alignment Mask. Experimental results on several benchmarks and corresponding ablation studies have demonstrated the effectiveness of the proposed method.*

## 1. Introduction

Unsupervised multi-class anomaly detection (AD) seek to classify and locate unusual regions of multi-class objects without relying on their labeled abnormal counterparts. It has widespread applications in single-modality and multimodal applications such as industrial quality inspection [1, 42], face anti-spoofing [18], etc. As emerging with great generative capability, diffusion models [11, 27] have shown remarkable potential in AD tasks. Treating AD tasks as con-

ditional generation problem, prior diffusion-based methods [9, 15, 22, 25, 33, 34, 39] assume they will generate normal counterparts confined to the learned normal distribution when conditioned on the given samples by condition injection [9, 25, 39] and in-sampling inversion [22, 33, 34]. The reconstruction errors of unknown test samples will provide measurable cues for anomaly identification.

For conditional diffusion-based AD methods, there exists a major issue called "identical shortcut" [15, 34, 35]. That is, abnormal information leaks from the conditioning input and accumulate through iterative denoising process, skewing generation results toward an anomalous data manifold. For condition injection, their condition modules are trained on normal samples and encouraged to absorb all details, thus leading them to be vulnerable to potential anomalous patterns in test samples. As for in-sampling inversion, subtle anomalies that cannot be fully masked by noise will be inadvertently amplified across iterative steps.

Recently, several studies have responded to this issue. For condition injection, VPOT [15] aims to filter out anomaly information in input conditional image, and learns prototypes as vague condition via optimal transport. To optimize in-sampling inversion, LafitE [34] utilizes memory bank and weighted integration to narrow the gap of initial state between conditional input image and learned normal distribution. However, the manner of prior methods to resolve this issue acts as double-edged sword, posing a trade-off challenge. As stated in [19], excessive suppression of anomaly information in conditional inputs will also erode critical details, and thus hinder from preserving consistency with normal regions. This limitation stems from that prior methods impose conditions as a time-invariant prior at the beginning or across the reverse progress.

The key to simultaneously address the "identical shortcut" issue and the trade-off challenge, is to design a proper time-variant condition mechanism. As observed by [13], condition has time-variant impacts. Condition added in early sampling have enormous influence on the generation trajectory, and the one added in relatively later timesteps has limited effect on overall direction.

\*Corresponding author.

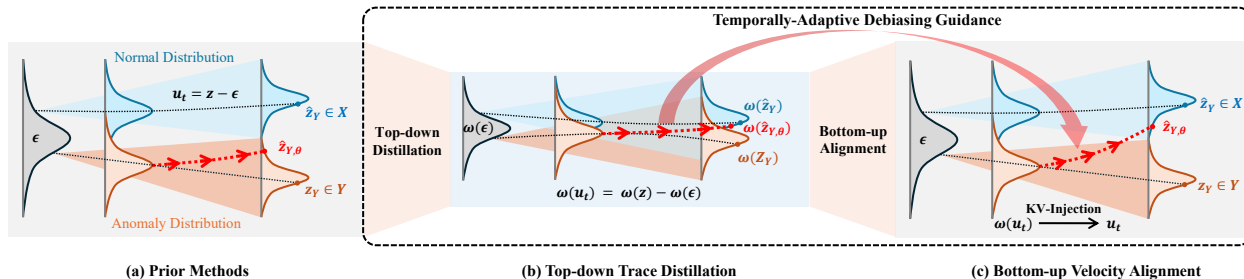


Figure 1. Illustration of the debiasing generation for normal distribution by the proposed Debiasing Trace Guidance when conditioned on anomaly distribution. Given a noise-data sample pair, top-down trace distillation constructs a low-dimension mapping subtrace with preserved correspondences, in which anomalous patterns are less distinguishable. Subsequently, bottom-up velocity alignment exploits latent features from low-dimension velocity field of the subtrace to guide a debiasing generation via KV-Injection Attention mechanism.

With this insight, we propose Debiasing Trace Guidance (DTG) framework based on Flow Matching [16], towards higher-precision unsupervised multi-class anomaly detection. Our core idea, depicted in Fig. 1, is to achieve debiasing generation by guiding it with time-varying velocity features from a distilled and thus anomaly-robust sub-trace.

We first propose Top-down Trace Distillation method to distill sub-traces in lower dimension from original generation traces via low-dimension mapping function  $\omega(\cdot)$ .  $\omega(\cdot)$  learns to extract general structures from  $z$  by learnable wavelet filters to filter out potential anomalies, and optimized by Skinhorn Distance to narrow the gap between filtered features from the same class in a batch. We predefine flow trace in straight form following [20]. In this way, velocity fields in different dimension are consistent when the mappings between noise and data samples are consistent. Therefore we properly design  $\omega(\cdot)$  to keep the mapping consistency. That is, given a pair of data sample  $z$  and noise  $\epsilon \sim \mathcal{N}(0, 1)$ ,  $\omega(\cdot)$  ensures the consistent surjection from  $\omega(z)$  to  $\omega(\epsilon) \sim \mathcal{N}(0, 1)$ . Then we propose Bottom-up Velocity Alignment method based on KV-Injection Attention with Alignment Mask for debiasing generation. Velocity features of sub-trace are firstly extract from UNet encoder at different timestep, and then injected into original high-dimension one via KV-Injection Attention. Then Alignment Mask forces the model to generate details of a specific region based on low-dimension global context from sub-trace velocity. Extensive experiments are conducted on several anomaly detection benchmarks, and the results demonstrate the effectiveness of the proposed DTG.

## 2. Related Work

**Unsupervised Anomaly Detection.** Unsupervised AD methods can be categorized by methodology into embedding-based and reconstruction-based methods, or by learning paradigm into single-class (one model for one class) and multi-class (one model for all classes) settings. Embedding-based methods [5, 6, 30] leverage learned rep-

resentations to identify deviations by similarity comparison. Reconstruction-based methods [23, 31, 35, 37, 40, 41], primarily built upon generative models, excel in handling complex anomaly patterns. The basic generative framework includes Auto-Encoder (AE) [35, 37, 40, 41], Variational Auto-Encoder (VAE) [23], and Generative Adversarial Net (GAN) [31]. The known "identical shortcut" problem for reconstruction-based methods is that, model could learn tricks so that anomalies are also restored well. Researchers have proposed different strategies to tackle this issue. JNLD [40], DRAEM [37], and OmniAL [41] focus on synthesis of pseudo-anomaly images. UniAD [35] prevents the model from learning the shortcut by masking neighbor tokens when calculating attention maps. With the same goal, HVQ-Trans [23] explores Vector Quantization to preserve normal patterns as discrete iconic prototypes.

**Diffusion Models for Anomaly Detection.** Diffusion models have represented powerful generative capability, and thus have been the mainstream [9, 15, 33, 34, 39] of reconstruction-based AD methods. Conditional generation of prior methods is mainly conducted through condition injection and in-sampling inversion. Condition injection [9, 25, 39] focuses on injecting test samples into the network through direct concatenation [25, 39], prompting modules or conditional adapters [9]. In-sampling inversion [22, 33] adds noise to the test sample till a certain time step and then recover from it via reverse process, which is applied in most of AD works to enhance pixel-level consistency with input samples. These two condition manners will both introduce leaking anomalous information. To tackle with this "identical shortcut" problem, prior methods focus on purify a time-invariant condition in terms of learnable injected condition or initial status of in-sampling inversion. VPOT [15] explores optimal transport to learn vague conditions as class-wise prototypes for condition injection, while Lafite [34] narrow the gap between initial states of learned distribution and testing images for in-sampling inversion. However, as stated in [19], they will fall into the trade-off

between the elimination of anomalies and the consistency with normal patterns. In this paper, we explore a novel condition mechanism to adaptively prioritizes anomaly suppression for the early generation timesteps, and turn to detail restoration in the later process.

### 3. Background

**Flow Matching.** In this paper, we represent diffusion generation process in context of flow matching [16]. We consider a generative model that mapping samples  $x_0$  from a tractable noise distribution  $p_0$  to samples  $x_1$  from data distribution  $p_1$ . The noise distribution  $p_0$  is usually set to be normal Gaussian distribution  $\mathcal{N}(0, I)$ , and  $p_1$  equals to normal data distribution  $q(X)$ . The mapping process is defined as an iterative denoising process in terms of ordinary differential equation (ODE) over  $t \in [0, 1]$ ,

$$dx_t = u(x_t, t)dt, \quad (1)$$

while generative model  $u(\cdot, t)$  defines a time-variant velocity field to generate a probability path  $p_t$  between  $p_0$  and  $p_1$ . Given  $x_0$ , the integration of velocity over time  $t$  is expected to form a flow trace  $\phi(x_0, x_1, t)$  that starts from  $x_0$  and ends at  $x_1 \in p_1$ , which consists of corresponding midpoints  $x_t$ . A specific form of flow trace  $\phi(x_0, x_1, t)$  determines the velocity field  $u(\cdot, t)$ . Thus, prior works tend to predefine  $\phi(x_0, x_1, t)$ . For the ease of linear Gaussian stochastic process, it is common to define  $\phi(x_0, x_1, t)$  as

$$x_t = \phi(x_0, x_1, t) = a_t x_0 + b_t x_1, \quad (2)$$

where  $a_t$  and  $b_t$  are hyperparameters that satisfy  $\phi(x_0, x_1, 0) = x_0$  and  $\phi(x_0, x_1, 1) = x_1$ . Prior works commonly set  $\{a_0 = 1, b_0 = 0\}$  and  $\{a_1 = 0, b_1 = 1\}$ .

A network  $G_\theta(\cdot, t)$  parameterized by  $\theta$  is used to predict  $u(\cdot, t)$ . Marginal vector field  $u(\cdot, t)$  is intractable and thus can not be directly used as training target. Instead, [16] exploits conditional vector field  $u(\cdot, t|x_0)$  defined on per sample and provides equivalent objective. Thus, the regression objective of  $G_\theta$  is defined as,

$$\mathcal{L}(\theta) = \mathbb{E}_{t, p_t(x_t|x_0), p(x_0)} \|u(x_t, t|x_0) - G_\theta(x_t, t)\|^2. \quad (3)$$

**Rectified Flow.** EDM [12] and Rectified Flow [20] define the flow trace  $\phi(x_0, x_1, t)$  as a straight path connecting noise sample  $x_0$  and data sample  $x_1$ ,

$$\phi(x_0, x_1, t) = (1 - t)x_0 + tx_1. \quad (4)$$

$\phi$  in Eq. 4 is utilized in this paper for the ease of velocity alignment, and accelerate generation within fewer steps.

### 4. Preliminaries

Let  $X$  and  $Y$  denote normal sample set and anomalous sample set respectively, containing multiple instance categories.

$q(X)$  and  $q(Y)$  are their corresponding data distribution. Let  $Z = \{z_X, z_Y\}$  be a mixed set where  $z_X \sim q(X)$  and  $z_Y \sim q(Y)$ . For every  $z \in Z$ , there exists  $\hat{z} \sim q(X)$  which keeps pixel-level consistency with  $z$ , except in suspect anomalous regions defined by  $\hat{m}$ , given by

$$\hat{z} = z - \hat{m}, \quad \hat{z} \sim q(X). \quad (5)$$

where deviation  $\hat{m}$  provides cues for anomaly detection and localization. In this paper, we focus on the debiasing generation to accurately reconstruct  $\hat{z} \in X$  from  $z \in Z$  under unsupervised limited access to normal samples  $z_X$ .

$\hat{z}$  is usually obtained via modeling conditional distribution  $q(\hat{z}|z)$  by parameterized  $p_\theta(\hat{z}|z)$ , where  $\theta$  denotes model weights. However, for our task,  $q(\hat{z}|z)$  can only be indirectly estimated since  $p(z)$  and  $\hat{z}$  are both intractable for unsupervised setting. Thus, we first model  $q(\hat{z}|z_X)$  by flow matching with condition injection, and then utilize it to indirectly estimate  $q(\hat{z}|z_Y)$  by in-sampling inversion. Given  $z \in Z$  with unknown label, suppose  $\hat{z}_t = \phi(\epsilon, \hat{z}, t)$  and  $z_t = \phi(\epsilon, z, t)$  are intermediate points obtained by the same  $\phi$  and  $\epsilon$ .  $\phi$  is set to straight flow trace in Eq. 4.  $\epsilon$  is sampled from  $\mathcal{N}(0, 1)$ , and  $\hat{z}_0 = \epsilon$ .  $\hat{z}$  is approximated as follows,

$$\begin{aligned} \hat{z} = \hat{z}_{t=1} &= \hat{z}_{t_0} + \int_{t_0}^1 u(\hat{z}_t, z[0], t)dt, \\ \hat{z}_{t_0} &:= (1 - t_0)\epsilon + t_0 z. \end{aligned} \quad (6)$$

The indirect estimation of  $q(\hat{z}|z)$  poses "identical shortcut" challenge and trade-off problem to both condition injection and in-sampling inversion, which can be addressed by our proposed trace distillation and velocity alignment.

## 5. Methodology

Overview of the proposed Debiasing Trace Guidance framework is shown in Fig. 2.

### 5.1. Top-down Trace Distillation

To provide guidance for debiasing velocity, we need first to construct a generation sub-trace robust to anomalies from the original trace. Then we finetune a siamese denoising unet on this distilled trace, and thus features at different timesteps can be extracted from the unet encoder as temporally-adaptive conditions. The difficulty of narrowing the gap between  $\hat{z}_{t_0}$  and  $z_{Y, t_0}$  is that there is no prior information about anomalous regions. Recent works [3, 32] apply average pooling to reserve the basic structure of images as generation guidance. The removal of high-frequency components will reduce the impacts of anomalies. Thus, for our task, it is suggested to find a low-dimension mapping function  $\omega(\cdot)$  for top-down distillation, i.e., approximating the flow trace  $\hat{z}_t$  in a low-dimension subspace.

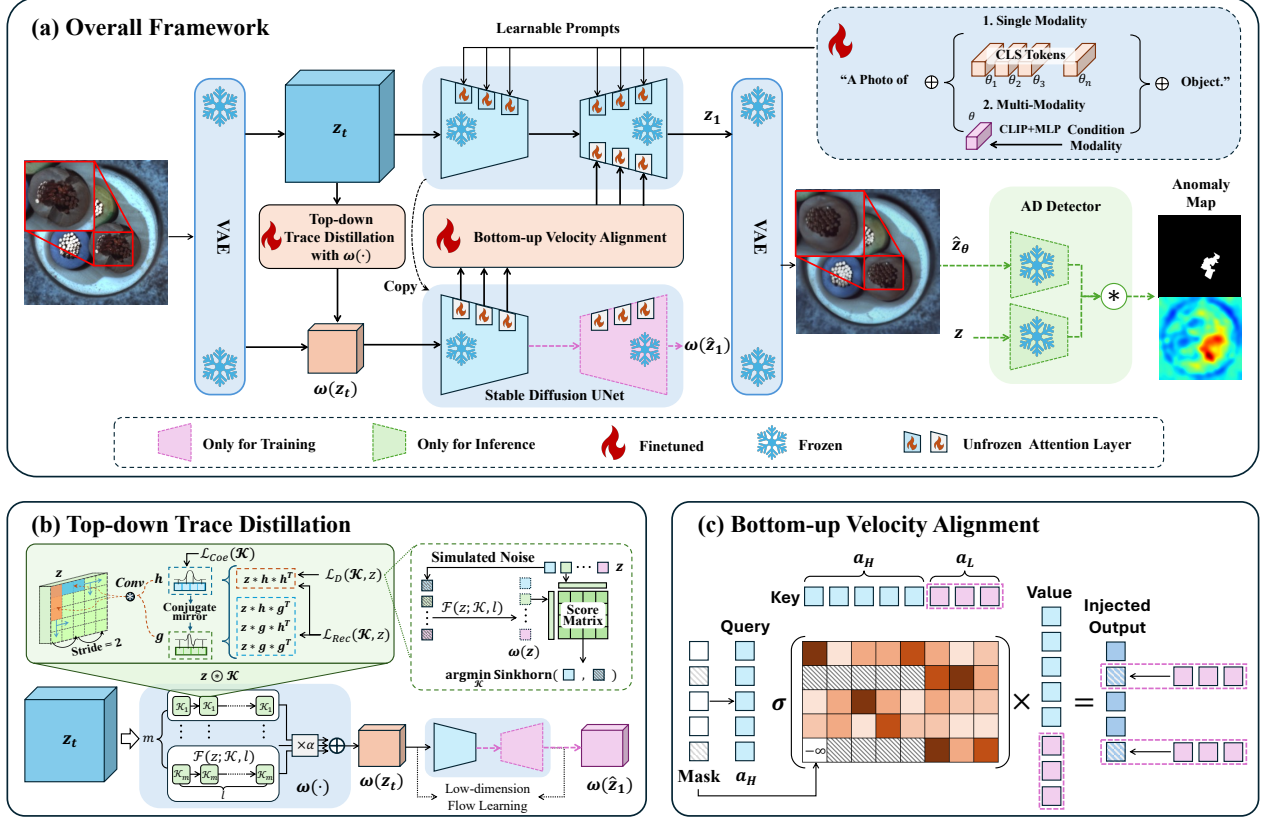


Figure 2. Overview framework of DTG, including Top-down Trace Distillation and Bottom-up Velocity Alignment.

**Guidelines for Low-dimension Mapping.** Since generation trace is predefined as straight path as in Eq. 4, velocities in different dimensions are consistent when the trace is consistent. Given the pair of  $z$  and  $\epsilon$ , the velocity field is tractable, i.e.,  $u_t = z - \epsilon$ .  $\omega(u_t)$  is expected to be a good approximation of  $\omega(\hat{u}_t)$  for  $\hat{z}$ , and we state that it should satisfy following three guidelines:

- $\omega(\cdot)$  should ensure the mapped low-dimension flow to be a surjection from data samples  $\omega(z[1])$  to normal Gaussian distribution  $\mathcal{N}(0, 1)$ , so that we can conduct bottom-up alignment to utilize this mapped  $\omega(u_t)$  for the guidance of the higher-dimension  $u_t$ .
- $\omega(\cdot)$  should learn to reserve the basic structure of  $z$  that also shared by  $\hat{z}$ .
- Additional conditions should be incorporated to avoid  $\omega(\cdot)$  collapsing to singular solutions like all-zero output.

The first guideline indicates that  $\omega(\cdot)$  should satisfy distributive property, i.e.,

$$\omega(u_t) = \omega(z[1] - \epsilon) = \omega(z[1]) - \omega(\epsilon), \quad (7)$$

which suggests to design  $\omega(\cdot)$  based on convolutional filter. Besides,  $\omega(\epsilon) \sim \mathcal{N}(0, 1)$  requires the sum of squares for kernel weights of  $\omega(\cdot)$  should equal to 1 for convolution.

**Learnable Convolutional Filter  $\omega(\cdot)$ .** Given above conditions,  $\omega$  is proposed as convolutional filter based on learnable wavelet decomposition under some elaborate constraints. Let define the lowpass filter used by  $\omega$  as  $h \in \mathbb{R}^{1 \times K}$ , which is implemented by 1D convolution kernel of length= $K$  and stride= $2$ . As stated in Corollary 1 and 2 of [2], there is  $\sum_k h[k] = \sqrt{2}$  for energy preservation. Besides,  $\|h\|_2^2 = 1$  to satisfy the orthogonality of the solution for basic recursion equation, as stated in Eq 3.13 of [2]. Notably, the unit  $\ell_2$  norm of wavelet lowpass kernel  $h$  just satisfies the 1-st guideline.

Then let us introduce the implementation of  $\omega(\cdot)$ . Given input data  $z$ , suppose  $\mathcal{F}(z; \mathcal{K}, l)$  denotes quad-tree 2D wavelet decomposition with kernel set  $\mathcal{K} = \{h, g\}$  and decomposition level  $l$ .  $g$  denotes a highpass filter which can be constructed from  $h$  as a conjugate mirror filter [8, 24],

$$g[k] = (-1)^n h[K - 1 - k]. \quad (8)$$

$\mathcal{F}(\cdot; \mathcal{K}, l)$  conducts separable 1D wavelet transform along vertical and horizontal axes for  $l$  levels, as follows,

$$\mathcal{F}(z; \mathcal{K}, l) = z \underset{i=1}{\overset{\circ}{\otimes}} \mathcal{K} \triangleq z \underset{l \text{ times}}{\otimes} \mathcal{K} \otimes \mathcal{K} \otimes \dots \otimes \mathcal{K}, \quad (9)$$

$$z \otimes \mathcal{K} = (z * h * h^T) \oplus (z * h * g^T) \oplus (z * g * h^T) \oplus (z * g * g^T),$$

where  $\otimes$  denotes decomposition operation,  $*$  represents convolution operation, and  $\oplus$  denotes concatenation operation along channel dimension.  $\mathcal{F}^{-1}(\cdot; \cdot)$  denotes wavelet reconstruction process with inverse kernel  $\mathcal{K}'$  satisfying  $z = \mathcal{F}^{-1}(\mathcal{F}(z; \mathcal{K}, l); \mathcal{K}', l)$ . The application of multiple  $\{h, g\}$  pairs like multiple experts will facilitate the capability to capture anomaly-robust features in low-dimension subspace. We rewrite  $\mathcal{K}$  as  $\mathcal{K} = [\mathcal{K}_1, \mathcal{K}_2, \dots, \mathcal{K}_m, \dots, \mathcal{K}_M]$  where  $M$  denotes the number of paired wavelet decomposition filters. In this way,  $\omega(\cdot; l)$  at level  $l$  is calculated by,

$$\omega(z) = \frac{1}{M} \sum_m \alpha_m \mathcal{F}(z; \mathcal{K}_m, l)[0], \quad (10)$$

where  $\mathcal{F}(\cdot; \cdot, \cdot)[0]$  means to return the first element along stacked channels, namely cascaded recursion of  $z * h * h^T$ .  $\alpha \in R^M$  denotes learnable score vector initialized with value 1.  $\omega(z)$  spatially compresses  $z$  with downsampling factor  $2^l$ , reserving necessary global structure from  $z$ .

**Optimization of Top-down Distillation.** We first guide  $\omega(\cdot)$  to learn to narrow the gap between  $\omega(\hat{z})$  and  $\omega(z)$ . As Perlin noise is ideal for mimicking anomalies due to its smooth, natural and controllable patterns, we follow [39] and mimic  $\hat{m}$  by Perlin noise  $m_p$  filled with textures from DTD dataset [4]. Thus, the distillation loss  $\mathcal{L}_D$  for training samples  $z_X \in X$  is

$$\mathcal{L}_D(\mathcal{K}, z) = \sum_i \mathcal{D}(\omega(z_{X,i}[1] + m_p), \omega(z_{X,i}[1])), \quad (11)$$

where  $\mathcal{D}$  denotes Skinhorn Distance. To avoid  $\mathcal{K}$  collapsing to singular solutions, loss  $\mathcal{L}_{Rec}$  for wavelet reconstruction is applied. It also forces  $\omega(z)$  to incorporate global context at level  $l$ , namely whole information of next level  $l + 1$ .

$$\mathcal{L}_{Rec}(\mathcal{K}, z) = \sum_i \|z_{X,i} - \mathcal{F}^{-1}(\mathcal{F}(z_{X,i}; \mathcal{K}, l); \mathcal{K}', l)\|_2^2. \quad (12)$$

To maintain surjection of distilled trace from data to  $\mathcal{N}(0, 1)$ , we impose constraint on kernel weights of  $h$  by coefficient loss  $\mathcal{L}_{Coef}$  as follows,

$$\mathcal{L}_{Coef}(\mathcal{K}) = \left\| \sum_k h[k] - \sqrt{2} \right\|_1 + \left\| \sum_k (h[k])^2 - 1 \right\|_1. \quad (13)$$

The final loss function  $\mathcal{L}_\omega$  to learn a proper  $\omega(\cdot)$  is,

$$\mathcal{L}_\omega = \mathcal{L}_D(\mathcal{K}, z) + \mathcal{L}_{Rec}(\mathcal{K}, z) + \mathcal{L}_{Coef}(\mathcal{K}). \quad (14)$$

## 5.2. Bottom-up Velocity Alignment

Based on distillation function  $\omega(\cdot)$  to approximate  $\omega(\hat{u}_t)$  by  $\omega(u_t)$ , the estimation of  $\hat{z}$  with  $u_t$  can be improved via aligning  $u_t$  with  $\omega(u_t)$ . Aligned  $u_t$  is expected to restore details from  $z_{t_0}$  and incorporate global context information from  $\omega(u_t)$  simultaneously. To achieve this, we propose KV-Injection Attention mechanism with Alignment Mask,

which forces the model to generate detailed information for a specific region based on low-dimensional global context.

**KV-Injection Attention.** Assume  $a_H \in \mathbb{R}^{N \times C}$  is serialized token feature of U-Net which is encoded from input data  $z$ , and  $a_L \in \mathbb{R}^{n \times C}$  is encoded from  $\omega(z)$ .  $N$  and  $n$  denote token length, and  $C$  denotes channel dimension. Based on attention modules in latent diffusion network, global structure from  $a_L$  is injected by concatenated with  $a_H$  for key and value of self-attention module. Query, key, and value are generated as follows,

$$\begin{cases} Q_H = \frac{a_H \cdot W_Q}{\sqrt{C}} \\ K_{HL} = (a_H \oplus a_L) \cdot W_K \\ V_{HL} = (a_H \oplus a_L) \cdot W_V, \end{cases} \quad (15)$$

where  $W_Q, W_K$ , and  $W_V \in \mathbb{R}^{C \times C}$  are all learnable parameters of the same shape,  $\cdot$  denotes matrix multiplication, and  $\oplus$  denotes spatial concatenation here.

**Alignment Mask for KV-Injection Attention.** To force  $u_t$  aligned  $\omega(u_t)$ , Alignment Mask  $\mathcal{M}(\cdot)$  is proposed to promote the transmission of global structure from  $a_L$  to  $a_H$ . The injected output  $a_{HL} \in \mathbb{R}^{N \times C}$  is

$$a_{HL} = \sigma(\mathcal{M}(Q_H \cdot K_{HL}^T)) \cdot V_{HL}, \quad (16)$$

where  $\sigma(\cdot)$  denotes Softmax function. To illustrate  $\mathcal{M}(\cdot)$ , we first define three types of indexes. Suppose  $\mathbf{i}_N$  denotes a index set  $\{i \mid i = 0, 1, 2, \dots, N\}$ , and  $\mathbf{i}_M$  denotes a subset of  $\mathbf{i}_N$ .  $\mathbf{i}_{N-M}$  denotes the rest valid indexes.  $\mathcal{M}(\cdot)$  applies random mask operation on attention scores of  $a_H$  by Eq. 17, and thus force the corresponding token of  $a_{HL}$  to be a interpolation of  $a_L$  by Eq. 18. Rewriting  $Q_H \cdot K_{HL}^T$  as  $M_{Attn}$ , Alignment Mask is represented by

$$\mathcal{M}(M_{Attn}) \rightarrow M_{Attn}[\mathbf{i}_M, 0 : N - 1] = -\infty. \quad (17)$$

where  $[\cdot]$  denotes index operation here. The assigning of negative infinity value leads the attention scores of  $a_H[\mathbf{i}]$  to be zero after Softmax function  $\sigma(\cdot)$ . Corresponding part  $a_{HL}[\mathbf{i}]$  of output tensor can be equivalently rewritten as

$$a_{HL}[\mathbf{i}_M] = \sigma(M_{Attn}[\mathbf{i}_M, N : N + n - 1]) \cdot a_L \cdot W_V, \quad (18)$$

where the obtained  $a_{HL}$  guarantees that, for randomly sampled  $\mathbf{i}_M$  elements of  $a_H$  which representing specific local areas in input image, it is forced to incorporate global context from  $a_L$  in training phase. In this way, the output velocity field  $u_t$  is aligned with  $\hat{u}_t$  for debiasing generation.

**Generation of Alignment Mask  $\mathbf{i}_M$ .** In training phase,  $\mathbf{i}_M$  are randomly sampled from  $\mathbf{i}_N$  by ratio  $r_M$ , whose size equals to rounded number of  $r_M N$ . In inference phase, the generation trace of anomalies tend to represent larger curvature than normal areas as in Fig. 3. Thus,  $\mathbf{i}_M$  are selected as the feature tokens with top- $r_M$  highest curvature, which

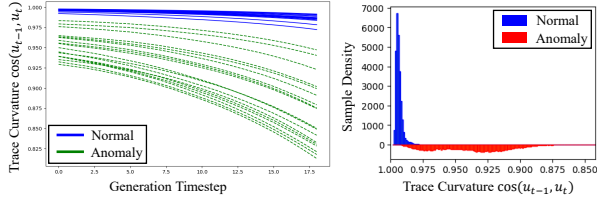


Figure 3. Analysis on generation trace curvatures of finetuned InstaFlow[21] for normal and anomalous samples in CASIA-SURF CeFA[18]. Trace curvature is represented by cosine value of average angle between predicted velocity vectors of adjacent timesteps. Generation are started from  $t = 0.1$  for 18 steps.

are more likely to be anomalous. The curvature is calculated on the fly as cosine value of average angle between predicted velocity vectors of past  $D$  adjacent timesteps, and lower value means higher curvature.  $D = 3$  for practice. Velocity vector denotes a latent token  $\in \mathbb{R}^{1 \times C}$  from  $a_{HL}$ .

## 6. Experiments

### 6.1. Datasets and Evaluation Metrics

**Datasets.** DTG is evaluated following unsupervised multi-class setting on three benchmarks, including single-modality datasets MVTeC-AD [1] and VisA [42], and cross-modal dataset CASIA-SURF CeFA [18] (CeFA), which jointly learns across different categories within one model.

MVTeC-AD dataset is a widely-used benchmark for industrial anomaly detection, containing over 5,354 image samples across 15 object and texture categories with pixel-level defect annotations. VisA dataset provides a larger scale and more diverse defect types, containing 10,821 images across 12 industrial categories, emphasizing complex structural anomalies as well as multi-object scenarios. CeFA dataset focuses on multi-modal cross-ethnicity face anti-spoofing, containing 1,607 subjects captured in RGB, infrared, and depth modalities with 2D/3D attack types and ethnicity labels. Notably, for cross-modal setting, we extract CLIP [26] token from RGB modal for condition injection at prompting module. The trace distillation and generation are conducted on Infrared modality.

**Evaluation Metrics.** Following prior works [15], we employ Area Under the Receiver Operating Characteristic curve (AUROC) for evaluation on MVTeC-Adversarial and VisA. Image-level  $AUROC_{cls}$  for detection and pixel-level  $AUROC_{seg}$  for localization are both applied. As for CeFA, it only has class label, and thus only detection task is conducted. Following [14, 18], we evaluate Average Classification Error Rate (ACER), Attack Presentation Classification Error Rate (APCER), and Bona Fide Presentation Classification Error Rate (BPCER) on CeFA dataset.

### 6.2. Implementation Details

**Network Settings.** We utilize InstaFlow [21] based on Stable Diffusion 1.5 [28] with rectified flow trace. The unets for low- and high-dimension traces share the same structure, and are finetuned by learnable prompts and unfrozen attention weights. For trace distillation, kernel length  $K$  of  $h$  is 6, level  $l$  is set to 2, and  $M = 8$  paired filters are utilized for  $\mathcal{K}$ . For velocity alignment, we extract low-dimension velocity features after the self-attention layer of the encoder. They are first embedded by a linear layer and then concatenated with high-dimension counterpart for key and value tokens. Learnable prompts are also inserted to two unets for finetuning. For single-modality settings, we add  $N_{CLS}$  prompting tokens  $h_c^*$  to learn class-related embeddings following textual inversion, which is inserted in the text template "a photo of  $h_c^*$  object.".  $N_{CLS}$  denotes the number of categories. Instead, for multi-modal settings, we stack one learnable linear layer with frozen CLIP [26] to extract prompting tokens from the conditional modality.

**Data Preprocess.** Input RGB images are resized to  $256 \times 256$  and augmented by random flip and rotation. Then they are encoded by VAE into  $32 \times 32 \times 3$ . For CeFA, the channel dimension of infrared images are duplicated to 3, and the values are normalized into  $[-1, 1]$ .

**Training Settings.** We divide training into three stages. Sampling  $z \in X$  and  $\epsilon$  pairs, we first learn the mapping function  $\omega$  with  $\mathcal{L}_\omega$ . Then  $\omega$  is fixed, and we learn the distilled traces by finetuning attention weights of a unet on pairs of  $\omega(z)$  and  $\omega(\epsilon)$ . After that, the encoder of unet is preserved and frozen as a conditional encoder. We finally finetune another unet for high-dimension velocity prediction with unfrozen attention weights and the guidance from the conditional encoder. We train for 400 epochs on a single NVIDIA L40 with a batch size of 16. AdamW is applied as optimizer with  $1e-5$  learning rate. The number of timesteps for training is set to 1000, consistent with [28].

**Inference Settings.** In inference phase, the number of generation timesteps is set to 20, and initial generation time  $t_0$  is set to 0.5. Thus, we generate  $\hat{z}$  within 10 timesteps. Mask ratio  $r_M$  for KV-Interpolation is set to 0.4. Following [33], anomaly score maps are calculated through both L1-Norm and perceptual-level distance, and processed by a Gaussian filter with kernel size = 5 for smooth.

### 6.3. Comparison with State-of-the-Art Methods

We conduct comparisons with state-of-the-art methods including feature embedded-based and reconstruction-based methods. Image-level and pixel-level accuracy metrics for anomaly detection are evaluated.

**Single-modality Anomaly Detection.** We first evaluate anomaly detection accuracy on MVTeC and VisA datasets. As shown in Table 1, our proposed DTG outperform previous works in both image-level and pixel-level accuracy

| Category            |             | Feature Embedding-based |                    |                    | Reconstruction-based |                    |                         |                          |                    |             |                    |                    |                    |
|---------------------|-------------|-------------------------|--------------------|--------------------|----------------------|--------------------|-------------------------|--------------------------|--------------------|-------------|--------------------|--------------------|--------------------|
|                     |             | MKD [30]                | RD [6]             | PatchCore [29]     | Non-Diffusion        |                    |                         | Diffusion                |                    |             |                    |                    |                    |
|                     |             |                         |                    |                    | DREAM [37]           | UniAD [35]         | InvAD [38]              | DiffAD [39]              | MDPS [33]          | DiAD [9]    | LafitE [34]        | VPOT [15]          | Ours               |
| Multi-class Setting |             | -                       | -                  | -                  | -                    | ✓                  | ✓                       | -                        | -                  | ✓           | ✓                  | ✓                  | ✓                  |
| Object              | Bottle      | 98.7 / 91.8             | 98.7 / 78.7        | <b>100</b> / 98.6  | 99.1 / 86.5          | 99.7 / 98.1        | <b>100</b> / 98.0       | <b>100</b> / <b>98.8</b> | <b>100</b> / 98.6  | 99.7 / 98.4 | <b>100</b> / 98.4  | <b>100</b> / 98.6  | <b>100</b> / 98.7  |
|                     | Cable       | 78.2 / 89.3             | 97.4 / 52.8        | <b>99.5</b> / 98.4 | 94.7 / 52.4          | 95.2 / 97.3        | 97.8 / <b>98.6</b>      | 94.6 / 96.8              | 98.3 / 95.8        | 94.8 / 96.8 | 97.2 / 97.3        | 97.8 / 98.1        | 99.4 / 98.2        |
|                     | Capsule     | 68.3 / 88.3             | 98.7 / 45.3        | 98.1 / 98.8        | 94.3 / 49.4          | 86.9 / 98.5        | <b>98.9</b> / 96.1      | 97.5 / 98.2              | 91.4 / 93.6        | 89.0 / 97.1 | 96.8 / <b>98.9</b> | 97.0 / 98.8        | 98.3 / 97.5        |
|                     | Hazelnut    | 97.1 / 91.2             | 98.9 / 61.2        | <b>100</b> / 98.7  | 99.7 / 92.9          | 99.8 / 98.1        | <b>100</b> / <b>100</b> | <b>100</b> / 99.4        | 99.8 / 98.6        | 99.5 / 98.3 | <b>100</b> / 98.4  | 99.9 / 98.7        | 98.1 / 97.6        |
|                     | Metal Nut   | 64.9 / 64.2             | 97.3 / 79.5        | <b>100</b> / 98.4  | 99.5 / 96.3          | 99.2 / 94.8        | 99.7 / 95.4             | <b>100</b> / <b>99.4</b> | 99.9 / 97.7        | 99.1 / 97.3 | 99.8 / 96.8        | 98.9 / 96.0        | 99.0 / 98.6        |
|                     | Pill        | 79.7 / 69.7             | 98.2 / 78.5        | 96.6 / 97.4        | 97.6 / 48.5          | 93.7 / 95.0        | <b>98.6</b> / 98.4      | 97.7 / 97.7              | 96.8 / <b>99.2</b> | 95.7 / 95.7 | 98.0 / 96.7        | 97.9 / 96.4        | 98.3 / 97.5        |
|                     | Screw       | 75.6 / 92.1             | <b>99.6</b> / 53.3 | 98.1 / 99.4        | 97.6 / 58.2          | 87.5 / 98.3        | 99.0 / 96.6             | 97.2 / 99.0              | 96.7 / 98.9        | 90.7 / 97.9 | 95.4 / <b>99.5</b> | 95.5 / 99.3        | 97.7 / 98.9        |
|                     | Toothbrush  | 75.3 / 88.9             | 99.1 / 50.5        | <b>100</b> / 98.7  | 98.1 / 44.7          | 94.2 / 98.4        | 98.9 / 96.9             | <b>100</b> / <b>99.2</b> | <b>100</b> / 98.8  | 99.7 / 99.0 | 92.9 / 98.8        | 94.6 / 98.8        | 98.8 / 99.1        |
|                     | Transistor  | 73.4 / 71.7             | 92.5 / 55.1        | <b>100</b> / 96.3  | 90.9 / 50.7          | 99.8 / 97.9        | 98.3 / 97.7             | 96.1 / 93.7              | <b>100</b> / 94.7  | 99.8 / 95.1 | 99.6 / 97.4        | 99.7 / <b>99.1</b> | 98.9 / 97.2        |
| Zipper              | 87.4 / 86.1 | 98.2 / 57.0             | 99.4 / 98.8        | 98.8 / 81.5        | 95.8 / 96.8          | 99.7 / <b>99.2</b> | <b>100</b> / 99.0       | <b>100</b> / 98.5        | 95.1 / 96.2        | 99.4 / 98.3 | 99.0 / 98.0        | 99.1 / 98.2        |                    |
| Texture             | Carpet      | 69.8 / 95.5             | 98.9 / 56.8        | 98.7 / 99.0        | 95.5 / 53.5          | 99.8 / 98.5        | 99.5 / <b>99.1</b>      | 98.3 / 98.1              | 99.6 / 94.4        | 99.4 / 98.6 | 99.8 / 98.7        | <b>100</b> / 98.8  | 98.7 / 98.4        |
|                     | Grid        | 83.8 / 82.3             | 99.3 / 49.6        | 98.2 / 98.7        | 99.7 / 65.7          | 98.2 / 96.5        | <b>100</b> / 96.3       | <b>100</b> / <b>99.7</b> | <b>100</b> / 99.4  | 98.5 / 96.6 | 99.9 / 98.5        | <b>100</b> / 99.4  | <b>100</b> / 99.4  |
|                     | Leather     | 93.6 / 96.7             | 99.4 / 47.7        | <b>100</b> / 99.3  | 98.6 / 75.3          | <b>100</b> / 98.8  | <b>100</b> / 99.3       | <b>100</b> / 99.1        | <b>100</b> / 99.5  | 99.8 / 98.8 | <b>100</b> / 99.2  | <b>100</b> / 99.2  | 99.6 / <b>99.7</b> |
|                     | Tile        | 89.5 / 85.3             | 95.6 / 53.2        | 98.7 / 95.6        | 99.2 / 92.3          | 99.3 / 91.8        | <b>100</b> / 94.2       | <b>100</b> / <b>99.4</b> | <b>100</b> / 96.4  | 96.8 / 92.4 | <b>100</b> / 93.2  | <b>100</b> / 94.5  | <b>100</b> / 99.1  |
|                     | Wood        | 93.9 / 80.5             | 95.3 / 48.8        | 99.2 / 95.0        | 96.4 / 77.7          | 98.6 / 93.2        | 97.7 / 95.1             | <b>100</b> / 96.7        | 99.1 / 95.7        | 99.7 / 93.3 | 99.7 / 94.3        | 98.2 / 95.3        | 99.3 / <b>98.3</b> |
| Mean                |             | 81.9 / 84.9             | 97.8 / 57.9        | <b>99.1</b> / 98.1 | 97.3 / 68.4          | 96.5 / 96.8        | 98.9 / 98.2             | 98.7 / 98.3              | 98.8 / 97.3        | 97.2 / 96.8 | 98.5 / 97.6        | 98.4 / 97.8        | 99.0 / <b>98.4</b> |

Table 1. Comparison on MVTeC-AD [1] with AUROC<sub>cls</sub>/AUROC<sub>seg</sub> metrics for anomaly detection / localization. Best results are bold.

| Category            |             | Non-Diffusion             |             |             |                           |                    |                    | Diffusion          |                           |                           |
|---------------------|-------------|---------------------------|-------------|-------------|---------------------------|--------------------|--------------------|--------------------|---------------------------|---------------------------|
|                     |             | JNLD [40]                 | OmniAL [41] | DRAEM [37]  | UniAD [35]                | HVQ-Trans [23]     | PatchCore [29]     | DiAD [9]           | VPOT [15]                 | Ours                      |
| Multi-class Setting |             | -                         | ✓           | -           | ✓                         | ✓                  | -                  | ✓                  | ✓                         | ✓                         |
| Complex Structure   | PCB1        | 82.9 / 98.0               | 77.7 / 97.6 | 83.9 / 94.0 | 95.4 / 99.3               | 96.7 / 99.4        | 98.5 / 94.3        | 88.1 / 98.7        | 98.2 / <b>99.6</b>        | <b>98.8</b> / 99.2        |
|                     | PCB2        | 79.1 / 95.0               | 81.0 / 93.9 | 81.7 / 94.1 | 93.6 / 97.8               | 93.4 / 98.0        | 97.3 / 89.2        | 91.4 / 95.2        | 97.5 / <b>98.8</b>        | <b>98.3</b> / 98.6        |
|                     | PCB3        | 90.1 / 98.5               | 88.1 / 94.7 | 87.7 / 94.1 | 88.6 / 98.3               | 92.0 / 98.3        | <b>97.9</b> / 90.9 | 86.2 / 96.7        | 94.5 / <b>98.7</b>        | 97.2 / 98.3               |
|                     | PCB4        | 96.2 / 97.5               | 95.3 / 97.1 | 87.1 / 72.3 | 99.4 / 97.9               | 99.5 / 97.7        | 99.6 / 90.1        | 99.6 / 97.0        | <b>99.9</b> / 97.8        | <b>99.9</b> / <b>99.2</b> |
| Multiple Instances  | Capsules    | <b>91.4</b> / <b>99.6</b> | 90.6 / 99.4 | 89.6 / 96.6 | 72.0 / 98.3               | 77.1 / 99.0        | 81.6 / 85.5        | 58.2 / 97.3        | 79.5 / 99.1               | 89.1 / 97.4               |
|                     | Candle      | 85.4 / 94.5               | 86.8 / 95.8 | 70.2 / 82.6 | 96.8 / 99.2               | 96.8 / 99.2        | <b>98.6</b> / 94.0 | 92.8 / 97.3        | 97.2 / <b>99.4</b>        | 98.4 / 99.1               |
|                     | Macaroni 1  | 90.5 / 93.3               | 92.6 / 98.6 | 68.6 / 89.8 | 92.2 / 99.3               | 93.1 / 99.4        | <b>97.5</b> / 95.4 | 85.7 / 94.1        | <b>97.5</b> / <b>99.6</b> | 90.5 / 99.3               |
| Single Instance     | Macaroni 2  | 71.3 / 92.1               | 75.2 / 97.9 | 60.3 / 83.2 | 85.9 / 98.0               | 86.2 / 98.5        | 78.1 / 94.4        | 62.5 / 93.6        | 85.7 / <b>99.0</b>        | <b>98.4</b> / 96.5        |
|                     | Cashew      | 82.5 / 94.1               | 88.6 / 95.0 | 67.3 / 68.5 | 92.4 / 98.7               | 94.9 / <b>99.2</b> | 97.3 / 94.5        | 91.5 / 90.9        | <b>99.0</b> / 98.0        | 98.2 / 97.9               |
|                     | Chewing Gum | 96.0 / 98.9               | 96.4 / 99.0 | 90.0 / 92.7 | <b>99.4</b> / <b>99.2</b> | <b>99.4</b> / 98.8 | 99.1 / 84.6        | 99.1 / 94.7        | 99.0 / 98.6               | 99.1 / 98.7               |
|                     | Fryum       | 91.9 / 90.0               | 94.6 / 92.1 | 86.2 / 83.2 | 89.8 / 97.7               | 90.4 / 97.7        | 96.2 / 85.3        | 89.8 / 97.6        | 92.0 / <b>98.6</b>        | <b>96.3</b> / 97.2        |
| Pipe Fryum          | 87.5 / 92.5 | 86.1 / 98.2               | 87.1 / 72.3 | 97.4 / 99.2 | 98.5 / <b>99.4</b>        | <b>99.8</b> / 95.7 | 96.2 / <b>99.4</b> | 98.8 / <b>99.4</b> | 98.6 / 99.1               |                           |
| Mean                |             | 87.1 / 95.2               | 87.8 / 96.6 | 80.5 / 87.0 | 91.9 / 98.6               | 93.2 / 98.7        | 95.1 / 91.2        | 86.8 / 96.0        | 94.2 / <b>98.9</b>        | <b>96.6</b> / 98.3        |

Table 2. Comparison on VisA [42] with AUROC<sub>cls</sub>/AUROC<sub>seg</sub> metrics for anomaly detection / localization. Best results are bold.

for MVTeC. DTG has taken a lead of 0.3 increment in AUROC<sub>cls</sub> and 0.1 increment in AUROC<sub>seg</sub>, when compared to prior best recorded method DiAD [9]. Notably, DTG has shown impressive performance for *Cable* and *Capsule*, suppressing DiAD by 4.8/0.8 in AUROC<sub>cls</sub> respectively. This can be attributed to temporally-adaptive debiasing guidance of DTG, which effectively filters out anomalies while be consistent with normal patterns. Accuracy comparison on VisA with more complex scenarios is shown in Table 2. DTG outperforms the second-best method VPOT [15] by 2.4 increment in AUROC<sub>cls</sub>. Especially for multiple instances, DTG gains average 4.1 increment in AUROC<sub>cls</sub>. It can be ascribed to that the guidance from preserved global structure by trace distillation facilitates multi-instance generation.

**Crossmodal Anomaly Detection.** We also evaluate the generalization of DTG via CeFA dataset with gray infrared maps and the guidance from RGB modality. As shown in Table 3, DTG gains the lowest error rate, and significantly outperforms the second-best method SpooTrace[14] by 52.0%/54.8% in ACER for Protocol 1/3 respectively.

**Visualization.** Fig. 4 demonstrates that DTG effectively eliminates potential anomalies and preserves normal details.

## 6.4. Ablation Study

Ablation studies are mainly conducted on MVTeC dataset. **Contribution of proposed components.** Table 4 demonstrates the effectiveness of the proposed top-down trace distillation and bottom-up velocity alignment. The baseline in first row is implemented by only finetuning the attention layers of InstaFlow [21]. Compared to baseline, DTG boosts accuracy by a 1.2 increment in AUROC<sub>cls</sub> and 0.8 increment AUROC<sub>seg</sub>. The fourth row demonstrates the significance of guideline restraint on  $\omega$ , as discussed in Sec. 5.1. Besides, the second row suggests that learnable promptings can facilitate generation quality.

**Manners of Trace Distillation.** For third row in Table 4, we set the same downsampling factor for average pooling with  $\omega(\cdot)$ . The result indicates that the main contribution of DTG lies in velocity guidance from a low-dimension trace, since a simple average pooling can boost 0.3 increment in AUROC<sub>cls</sub>. The proposed trace distillation further improve the debiasing guidance with 0.5 increment in AUROC<sub>cls</sub>.

**Decomposition Level for  $\omega$ .** We further dive into the decomposition level for  $\omega$ . Larger level  $l$  means to extract more global structure information from input condition, and will provide less details. Since the resolution of encoded la-

| Method         | Protocols       |          |            |           |          |          |                     |          |            |
|----------------|-----------------|----------|------------|-----------|----------|----------|---------------------|----------|------------|
|                | Cross-Ethnicity |          |            | Cross-PAI |          |          | Cross-Ethnicity&PAI |          |            |
|                | ACER(%)         | APCER(%) | BPCER(%)   | ACER(%)   | APCER(%) | BPCER(%) | ACER(%)             | APCER(%) | BPCER(%)   |
| PSMM-Net[18]   | 3.5             | 2.4      | 4.6        | 5.4       | 7.7      | 3.1      | 6.7                 | 7.8      | 5.5        |
| MC-PixBis[10]  | 15.1            | 1.4      | 28.7       | 5.9       | 10.6     | 1.2      | 15.9                | 19.0     | 12.8       |
| FaceBagNet[17] | 17.4            | 2.1      | 32.7       | 7.9       | 15.5     | 0.9      | 26.7                | 37.9     | 15.4       |
| CDCN[36]       | 6.8             | 0.0      | 13.6       | 1.2       | 0.0      | 2.5      | 9.7                 | 0.5      | 19.0       |
| CMFL[7]        | 13.5            | 3.7      | 23.4       | 2.2       | 3.6      | 0.9      | 15.2                | 11.6     | 19.0       |
| SpoofTrace[14] | 5.0             | 2.1      | 7.9        | 0.9       | 0.2      | 1.5      | 6.2                 | 1.3      | 11.1       |
| Ours           | <b>2.4</b>      | 1.2      | <b>3.5</b> | 2.0       | 0.6      | 3.4      | <b>2.8</b>          | 2.3      | <b>3.3</b> |

Table 3. Comparison on CeFA [18] dataset for anomaly detection. Average ACER/APCER/BPCER metrics are evaluated on three protocols from CeFA. Lower metric value denotes better face anti-spoofing accuracy.

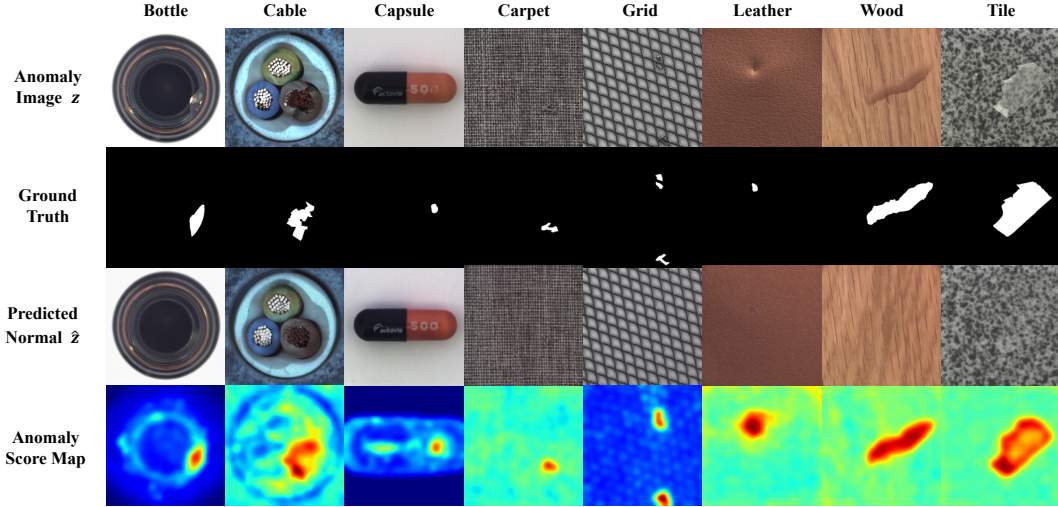


Figure 4. Visualization of reconstructed  $\hat{z}$  conditioned  $z$  by DTG and corresponding anomaly score map on MVTeC dataset.

| Condition Distillation |                    |                     | Condition Injection Manner |                    | Accuracy    |
|------------------------|--------------------|---------------------|----------------------------|--------------------|-------------|
| Avg Pooling            | Trace Distillation | $L_{Rec} + L_{Coe}$ | Prompting                  | Velocity Alignment |             |
| -                      | -                  | -                   | -                          | -                  | 97.8 / 97.6 |
| -                      | -                  | -                   | ✓                          | -                  | 98.2 / 97.8 |
| ✓                      | -                  | -                   | ✓                          | ✓                  | 98.5 / 98.1 |
| -                      | ✓                  | -                   | ✓                          | ✓                  | 97.1 / 96.8 |
| -                      | ✓                  | ✓                   | ✓                          | ✓                  | 99.0 / 98.4 |

Table 4. Ablation study on proposed components of DTG with  $AUROC_{cls}/AUROC_{seg}$  metrics for MVTeC. ✓ means being utilized.

| Level of $\omega$              | l=1         | l=2         | l=4         |
|--------------------------------|-------------|-------------|-------------|
| Accuracy                       | 98.8 / 98.4 | 99.0 / 98.4 | 98.5 / 98.2 |
| Mask ratio $r_M$ for inference | 0.2         | 0.4         | 0.6         |
| Accuracy                       | 98.8 / 98.3 | 99.0 / 98.4 | 98.3 / 97.9 |

Table 5. Ablation studies on decomposition level  $l$  of  $\omega(\cdot)$  and mask ratio  $r_M$  with  $AUROC_{cls}/AUROC_{seg}$  metrics for MVTeC.

tents by VAE is  $32 \times 32$  for  $256 \times 256$  input, excessive downsampling factor is also harmful. As shown in Table 5,  $l = 2$  is a proper selection.

**Alignment Extent via Mask Ratio  $r_M$ .** Mask ratio  $r_M$  adjust the extent to which the high-dimension velocity is aligned with the lower-dimension counterpart. The sec-

ond row of Table 5 suggests that,  $r_M$  lower than a specific threshold will facilitate the debiasing guidance from a robust distilled velocity field to the original one. And thus we choose  $r_M = 0.4$  for practice.

## 7. Conclusion

In this paper, we propose a novel Debiasing Trace Guidance (DTG) framework, which introduces debiasing generation to address "identical shortcut" issue towards higher-precision unsupervised multi-class anomaly detection. DTG resolves trade-off between suppressing anomalies and preserving normal details. First, Top-down Trace Distillation is proposed to construct a sub-trace from the main trace that preserves global structural information robust to anomalies and the correspondence of noise-sample pairs. Then, DTG proposes Bottom-up Velocity Alignment to inject low-dimensional velocity features into main generation process via KV-Injection Attention mechanism and Alignment Mask. In this way, DTG achieves debiasing generation by elimination of anomalies and preservation of normal details. Extensive experimental results demonstrate that DTG achieves state-of-the-art performance.

## Acknowledgements

This work is supported in part by National Natural Science Foundation of China under Grant U2441244, and Zhejiang Provincial Natural Science Foundation of China under Grant LZ24F030006.

## References

- [1] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattler, and Carsten Steger. The MVTEC Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021. 1, 6, 7
- [2] C.S. Burrus, R.A. Gopinath, and H. Guo. *Introduction to Wavelets and Wavelet Transforms: A Primer*. 1998. 4
- [3] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv:2207.12598*, 2021. 3
- [4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5
- [5] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 475–489, 2021. 2
- [6] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9737–9746, 2022. 2, 7
- [7] Anjith George and Sébastien Marcel. Cross modal focal loss for rgb-d face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7882–7891, 2021. 8
- [8] Wooseok Ha, Chandan Singh, Francois Lanusse, Srigokul Upadhyayula, and Bin Yu. Adaptive wavelet distillation from neural networks through interpretations. *Advances in Neural Information Processing Systems*, 34, 2021. 4
- [9] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A Diffusion-Based Framework for Multi-Class Anomaly Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):8472–8480, 2024. 1, 2, 7
- [10] Guillaume Heusch, Anjith George, David Geissbühler, Zohreh Mostaani, and Sébastien Marcel. Deep models and shortwave infrared information to detect face presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):399–409, 2020. 8
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv:2006.11239*, 2020. 1
- [12] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. *arXiv:2206.00364*, 2022. 3
- [13] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models, 2024. 1
- [14] Kaicheng Li, Hongyu Yang, Binghui Chen, Pengyu Li, Biao Wang, and Di Huang. Learning Polysemantic Spoof Trace: A Multi-Modal Disentanglement Network for Face Anti-spoofing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):1351–1359, 2023. 6, 7, 8
- [15] Yuxin Li, Yaoxuan Feng, Bo Chen, Wenchao Chen, Yubiao Wang, Xinyue Hu, Baolin Sun, Chunhui Qu, and Mingyuan Zhou. Vague Prototype-Oriented Diffusion Model for Multi-Class Anomaly Detection. In *Proceedings of the 41st International Conference on Machine Learning*, pages 27771–27790, 2024. 1, 2, 6, 7
- [16] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv:2210.02747*, 2023. 2, 3
- [17] Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, et al. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019. 8
- [18] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z. Li. CASIA-SURF CeFA: A Benchmark for Multi-Modal Cross-Ethnicity Face Anti-Spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1179–1187, 2021. 1, 6, 8
- [19] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-Measurable Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12147–12156, 2023. 1, 2
- [20] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv:2209.03003*, 2022. 2, 3
- [21] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflo: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024. 6, 7
- [22] Fanbin Lu, Xufeng Yao, Chi-Wing Fu, and Jiaya Jia. Removing anomalies as noises for industrial defect localization. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16120–16129, 2023. 1, 2
- [23] Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. In *Advances in Neural Information Processing Systems*, pages 8487–8500, 2023. 2, 7
- [24] Gabriel Michau, Gaetan Frusque, and Olga Fink. Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series. *Proceedings of the National Academy of Sciences*, 119(8):e2106598119, 2022. 4
- [25] Arian Mousakhan, Thomas Brox, and Jawad Tayyub. Anomaly detection with conditioned denoising diffusion models, 2023. 1, 2

- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. [6](#)
- [27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. [1](#)
- [28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. [6](#)
- [29] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection, 2021. [7](#)
- [30] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14902–14912, 2021. [2, 7](#)
- [31] Jouwon Song, Kyeongbo Kong, Ye-In Park, Seong-Gyun Kim, and Suk-Ju Kang. Anomaly segmentation network using self-supervised learning. In *AAAI 2022 Workshop on AI for Design and Manufacturing (ADAM)*, 2021. [2](#)
- [32] Zhixin Wang, Ziyang Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. DR2: Diffusion-Based Robust Degradation Remover for Blind Face Restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1704–1713, 2023. [3](#)
- [33] Di Wu, Shicai Fan, Xue Zhou, Li Yu, Yuzhong Deng, Jianxiao Zou, and Baihong Lin. Unsupervised Anomaly Detection via Masked Diffusion Posterior Sampling, 2024. [1, 2, 6, 7](#)
- [34] Haonan Yin, Guanlong Jiao, Qianhui Wu, Borje F. Karlsson, Biqing Huang, and Chin Yew Lin. Lafite: Latent diffusion model with feature editing for unsupervised multi-class anomaly detection, 2023. [1, 2, 7](#)
- [35] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Advances in Neural Information Processing Systems*, 2022. [1, 2, 7](#)
- [36] Zitong Yu, Yunxiao Qin, Xiaobai Li, Zezheng Wang, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Multi-modal face anti-spoofing based on central difference networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 650–651, 2020. [8](#)
- [37] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. DRAEM - a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8330–8339, 2021. [2, 7](#)
- [38] Jiangning Zhang, Chengjie Wang, Xiangtai Li, Guanzhong Tian, Zhucun Xue, Yong Liu, Guansong Pang, and Dacheng Tao. Learning feature inversion for multi-class anomaly detection under general-purpose coco-ad benchmark. (arXiv:2404.10760), 2024. [7](#)
- [39] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised Surface Anomaly Detection with Diffusion Probabilistic Model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6759–6768, 2023. [1, 2, 5, 7](#)
- [40] Ying Zhao. Just noticeable learning for unsupervised anomaly localization and detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 01–06, 2022. [2, 7](#)
- [41] Ying Zhao. Omnial: A unified cnn framework for unsupervised anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3924–3933, 2023. [2, 7](#)
- [42] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-Difference Self-supervised Pre-training for Anomaly Detection and Segmentation. In *European Conference on Computer Vision*, pages 392–408, 2022. [1, 6, 7](#)