Paper Review

Cheating Depth: Enhancing 3D Surface Anomaly Detection via Depth Simulation

YeongHyeon Park

Department of Electrical and Computer Engineering

SungKyunKwan University

Cheating Depth: Enhancing 3D Surface Anomaly Detection via Depth Simulation

Vitjan Zavrtanik Matej Kristan Danijel Skočaj
Faculty of Computer and Information Science, University of Ljubljana
{vitjan.zavrtanik, matej.kristan, danijel.skocaj}@fri.uni-lj.si

Abstract

RGB-based surface anomaly detection methods have advanced significantly. However, certain surface anomalies remain practically invisible in RGB alone, necessitating the incorporation of 3D information. Existing approaches that employ point-cloud backbones suffer from suboptimal representations and reduced applicability due to slow processing. Re-training RGB backbones, designed for faster dense input processing, on industrial depth datasets is hindered by the limited availability of sufficiently large datasets. We make several contributions to address these challenges. (i) We propose a novel Depth-Aware Discrete Autoencoder (DADA) architecture, that enables learning a general discrete latent space that jointly models RGB and 3D data for 3D surface anomaly detection. (ii) We tackle the lack of diverse industrial depth datasets by introducing a simulation process for learning informative depth features in the depth encoder. (iii) We propose a new surface anomaly detection method 3DSR, which outperforms all existing state-of-the-art on the challenging MVTec3D anomaly detection benchmark, both in terms of accuracy and processing speed. The experimental results validate the effectiveness and efficiency of our approach, highlighting the potential of utilizing depth information for improved surface anomaly detection. Code is available at: https://github.com/VitjanZ/3DSR

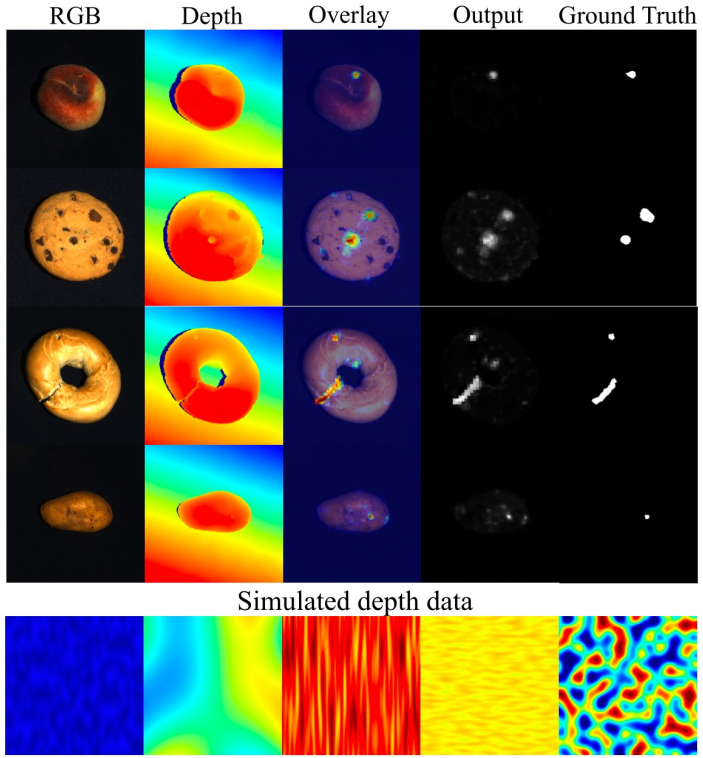


Figure 1. Certain anomalies are practically imperceptible in RGB, requiring depth for precise detection. Parameterized generative model yields images sufficiently describing depth statistics for training general depth-reconstruction backbones.

Vitjan Zavrtanik

University of Ljubljana
Verified email at fri.uni-lj.si

Computer vision anomaly detection

TITLE	CITED BY	YEAR
Reconstruction by inpainting for visual anomaly detection V Zavrtanik, M Kristan, D Skočaj Pattern Recognition 112, 107706	259	2021
Draem-a discriminatively trained reconstruction embedding for surface anomaly detection V Zavrtanik, M Kristan, D Skočaj Proceedings of the IEEE/CVF International Conference on Computer Vision …	224	2021
Dsr–a dual subspace re-projection network for surface anomaly detection V Zavrtanik, M Kristan, D Skočaj European conference on computer vision, 539-554	21	2022

Cheating Depth

Summaries

Motivation and solution

- Motivation: Some surface anomalies are practically invisible in RGB space
- Solution: Integration of 3D information with RGB image

> Alternatively, RGB backbones could be re-trained on industrial depth datasets, but the current industrial depth datasets are too small to efficiently train the large backbones. Recent work DSR [22] proposed utilizing Vector-quantized autoencoders (VQVAE) [13], which learn only a fixed number of discrete latent representation vectors, thus potentially enable learning from smaller datasets. Nevertheless, our experience shows that training DSR on the available industrial depth dataset MVTec3D [2] leads to suboptimal results, indicating that the existing data is too small even for the representation-efficient VQVAEs. Advances in RGB+3D surface anomaly detection are thus hindered by the lack of sufficiently large datasets that would enable pre-training general depth backbones and allow development of methods fast enough for practical applications.

Contributions

- Achieve a high performance with small scale dataset
  - A fixed number of discrete latent representations of VQVAE enables training with a small-scaled dataset
  - Simulation process for learning informative depth-feature (training with synthetic data)
- 3DSR* to exploit joint modality information of RGB and depth
  - 3DSR Includes one DADA encoder, two latent decoders
  - DADA** learns joint representations of RGB and 3D data
  - State-of-the-art anomaly detection performance and processing time (w/ RTX A4500)

3D sensor

Availability: Development kit available for qualified customers.
Price: EUR 7500.

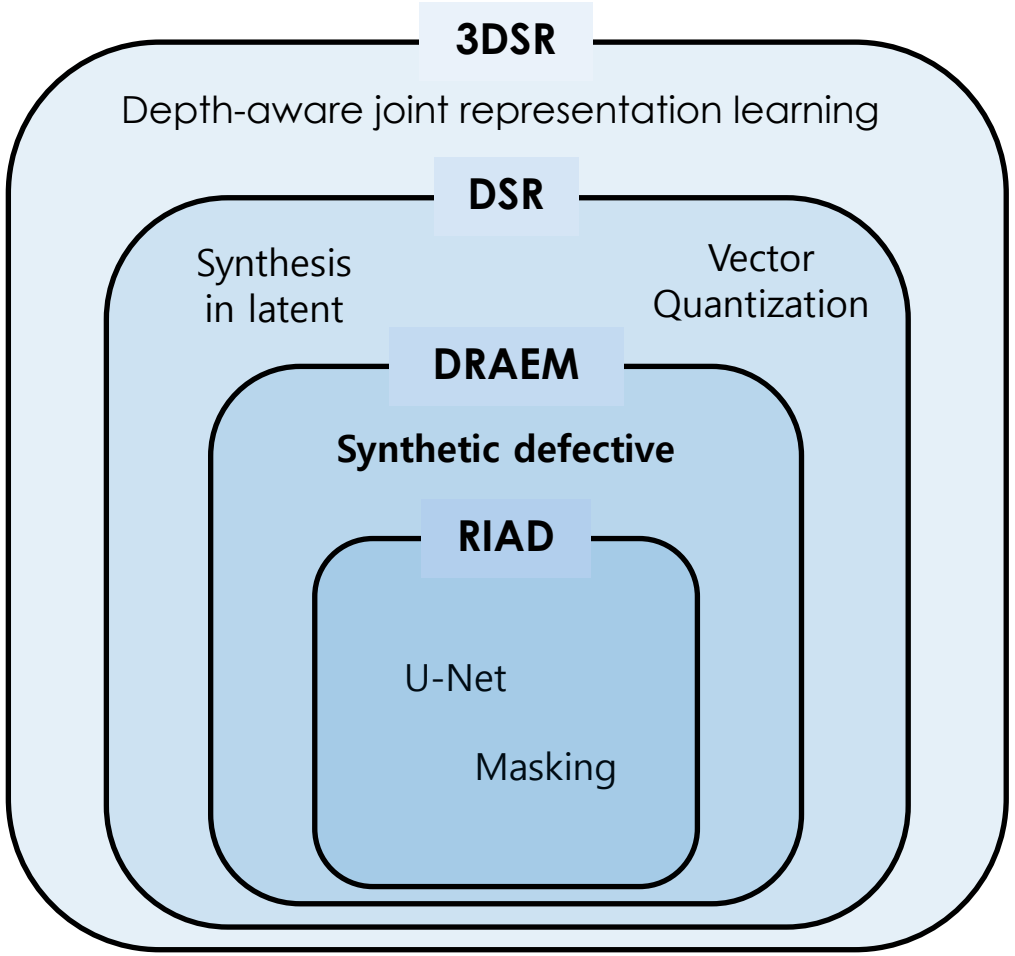
Remaining issues

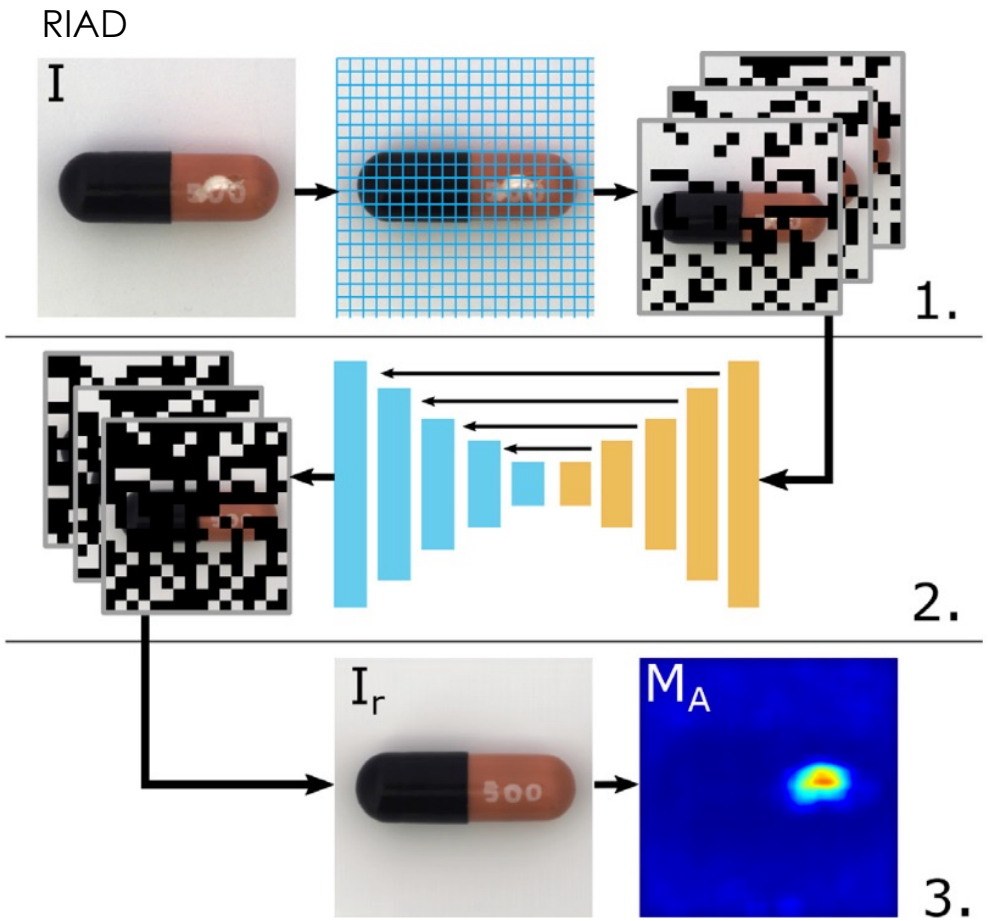
- What if we/who cannot afford expensive 3D sensors?
- Why unrelated ImageNet-random simulated depth pair training is effective?

* 3DSR: 3D Dual Subspace Reprojection
** DADA: Depth-Aware Discrete Autoencoder

Vitjan Zavrtanik et al. "Reconstruction by inpainting for visual anomaly detection." Pattern Recognition. 2021.
Vitjan Zavrtanik et al. "DRAEM-a discriminatively trained reconstruction embedding for surface anomaly detection." ICCV. 2021.
Vitjan Zavrtanik et al. "DSR–a dual subspace re-projection network for surface anomaly detection." ECCV, 2022.
Vitjan Zavrtanik et al. "Cheating Depth: Enhancing 3D Surface Anomaly Detection via Depth Simulation." WACV. 2024.

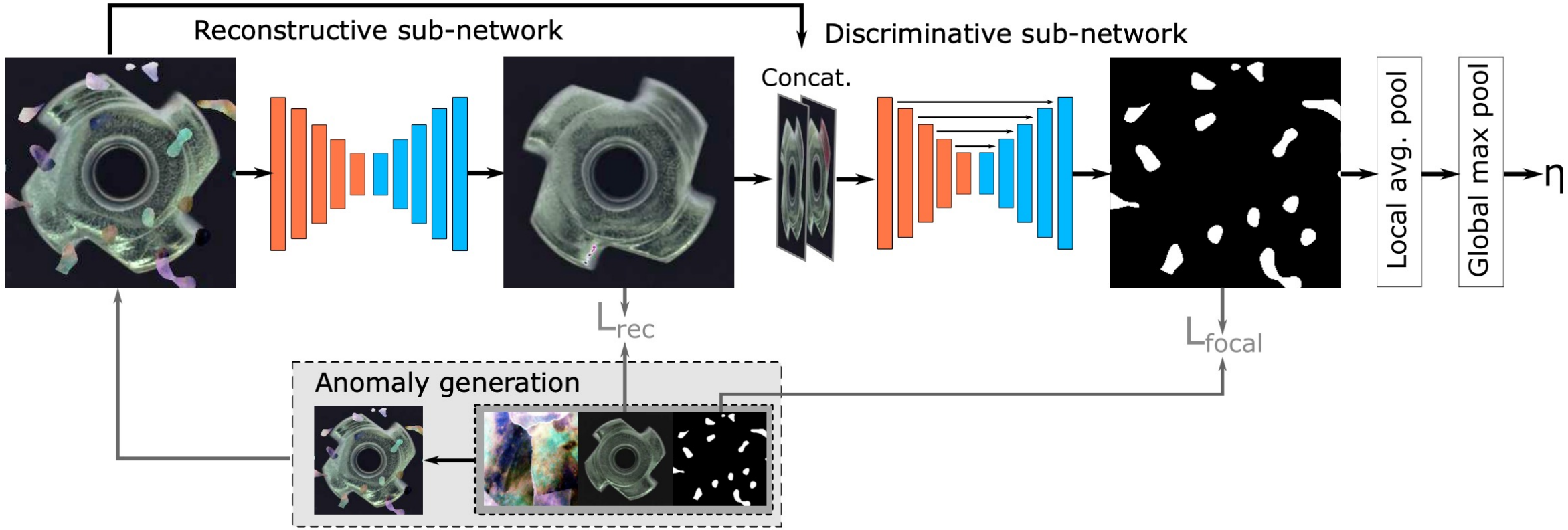
Vitjan Universe





Vitjan Zavrtanik et al. "DRAEM-a discriminatively trained reconstruction embedding for surface anomaly detection." ICCV. 2021.

DRAEM



DSR: Dual Surface Reprojection

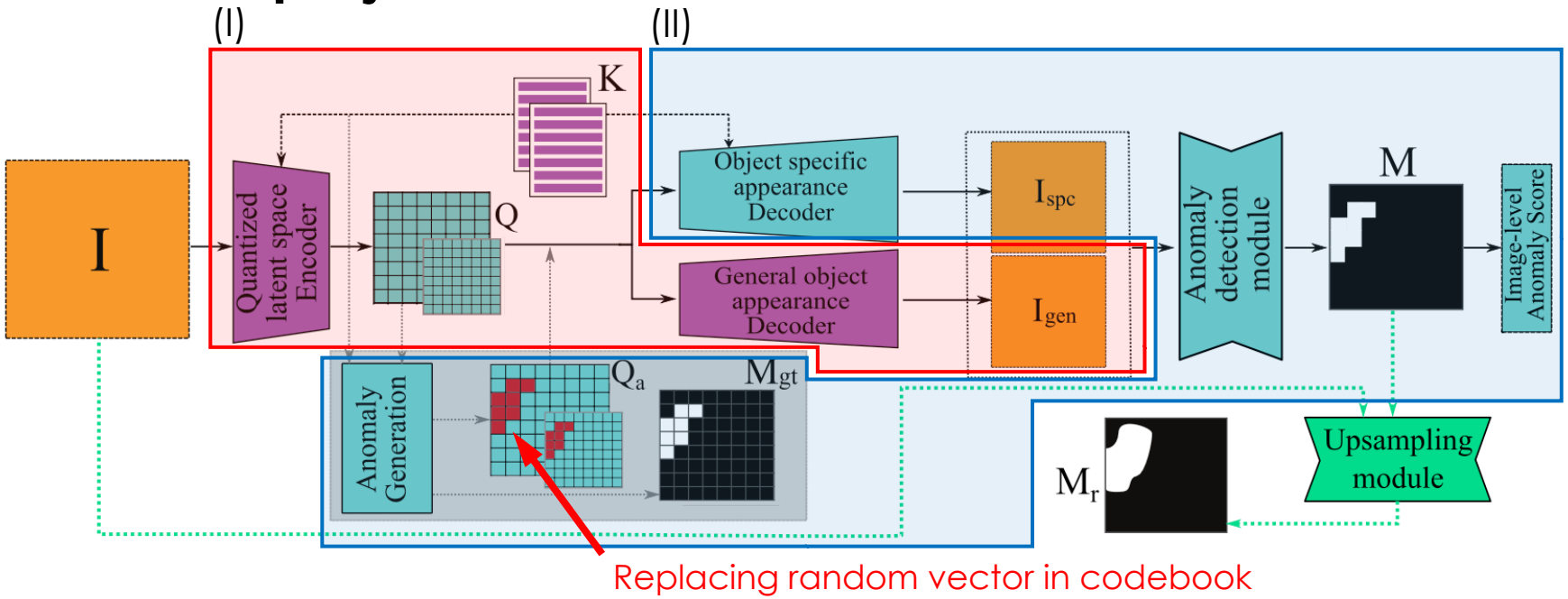


Fig. 2. The DSR architecture. During training, the non-anomalous image quantized feature maps (Q_{hi}, Q_{lo}) are replaced by the anomaly augmented feature maps $(Q_{a,hi}, Q_{a,lo})$ generated by the latent space sampling procedure (shaded block). The pathway marked with green arrows are used when training the Upsampling module with simulated smudges and at inference.

I. Training stage 1. representation learning of ImageNet dataset
   - Learns discrete VQ representations

II. Training stage 2. representation learning…
   - To translate synthetic anomalous samples into a normal form
   - To segment potentially anomalous regions.

Motivations

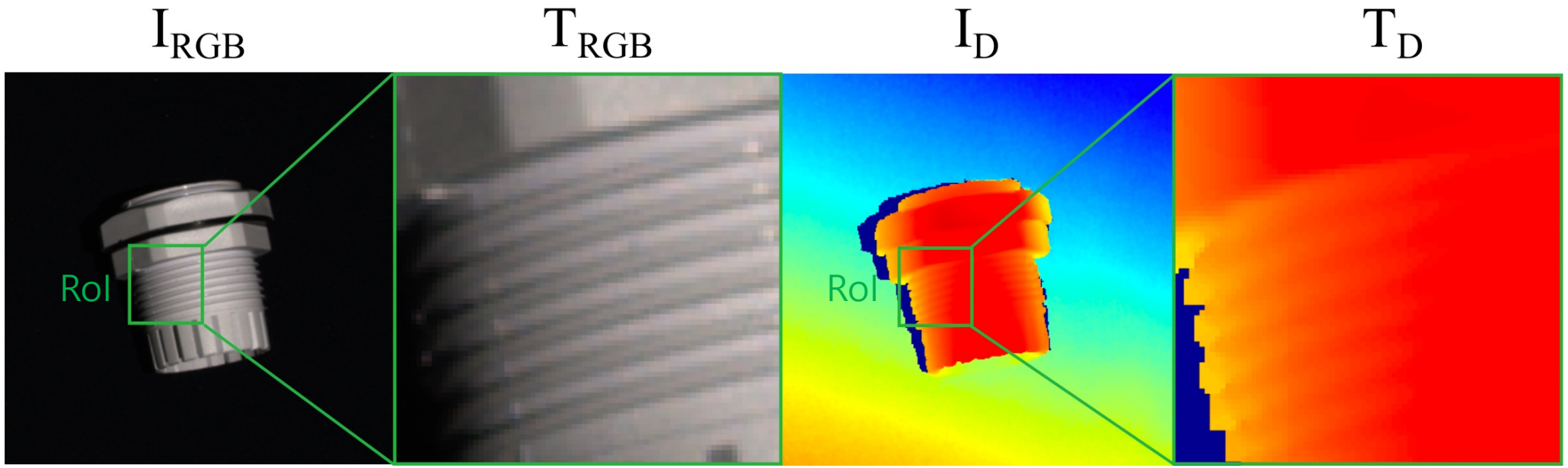


Figure 2. Example of cable gland from the MVTec3D dataset [2].

- A RoI in RGB shows high variation (high local gradients) due to shadows
- In contrast, a RoI in depth shows low variation → depth variations can be informative

Paul Bergmann et al. "The MVTec 3D-AD dataset for unsupervised 3d anomaly detection and localization." arXiv. 2021.
Zivid One+ Medium, https://www.zivid.com/zivid-one-plus-medium-3d-camera

MVTec 3D-AD

10 subtasks achieved by high-resolution industrial 3D sensors

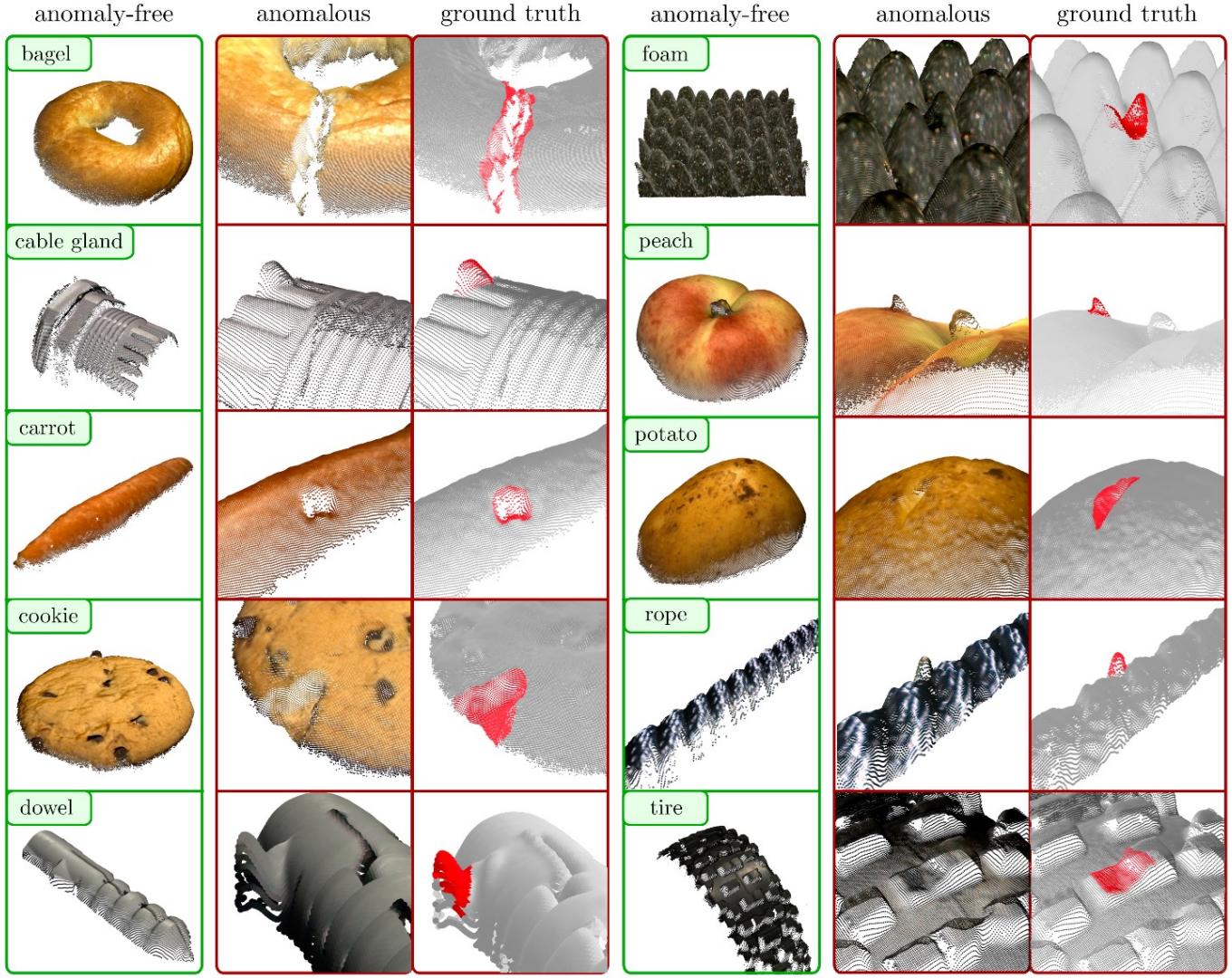
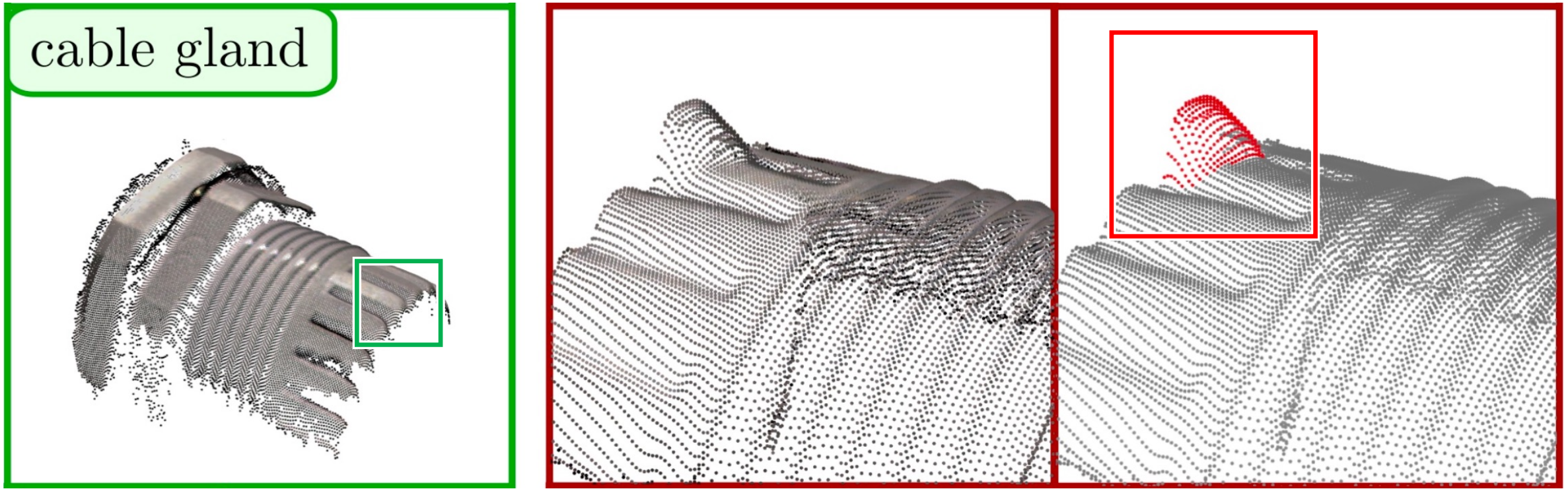


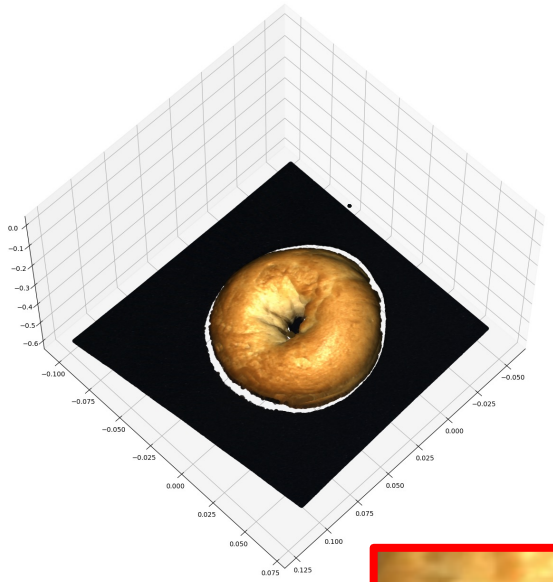
Figure 2: Examples for all 10 dataset categories of the MVTec 3D-AD dataset. For each category, the left column shows an anomaly-free point cloud with RGB values projected onto it. The second column shows a close-up view of an anomalous test sample. Anomalous points are highlighted in the third column in red. Note that the background planes were removed for better visibility.

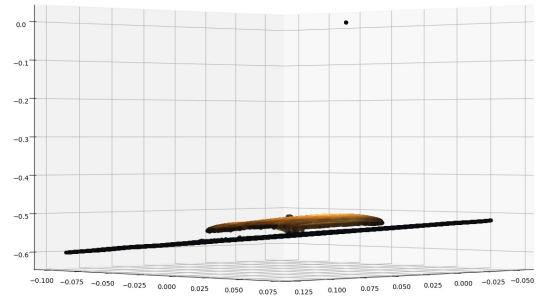
MVTec 3D-AD



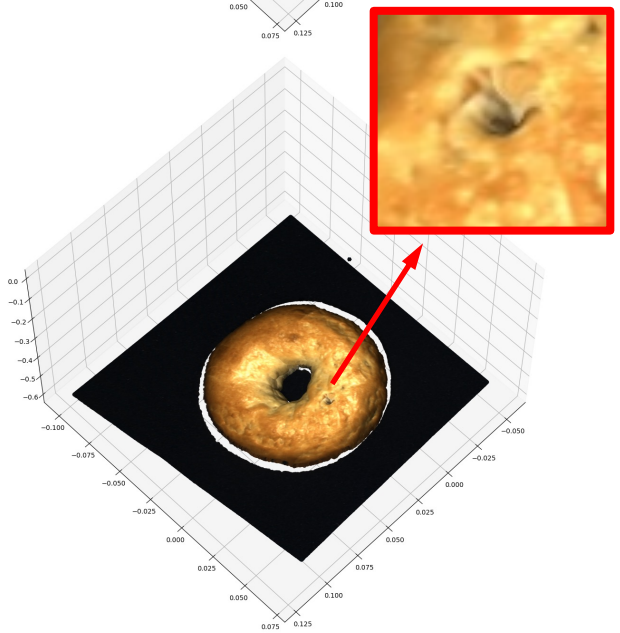
Samples

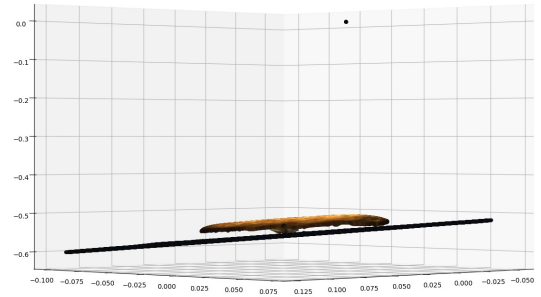
Good





Anomalous





3DSR: 3D Dual Subspace Reprojection

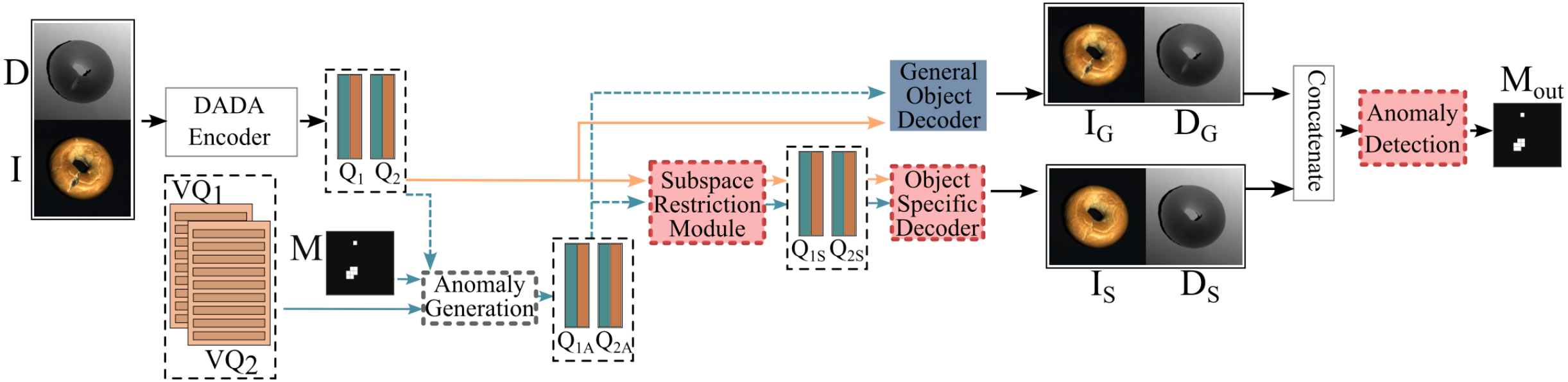


Figure 5. Training process of 3DSR. The red boxes are only used for training.

3DSR: 3D Dual Subspace Reprojection

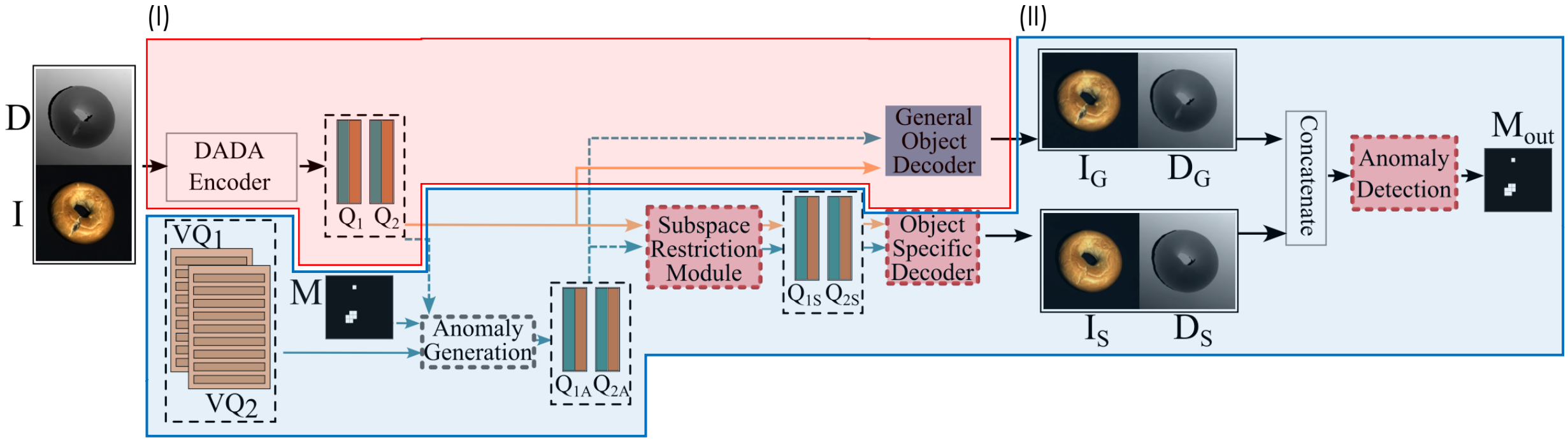


Figure 5. Training process of 3DSR. The red boxes are only used for training.

I. Training stage 1. representation learning of ImageNet dataset
   - Learns joint modality discrete VQ representations by DADA

II. Training stage 2. representation learning…
   - To translate synthetic anomalous samples into a normal form
   - To segment potentially anomalous regions.

DADA: Depth-Aware Discrete Autoencoder

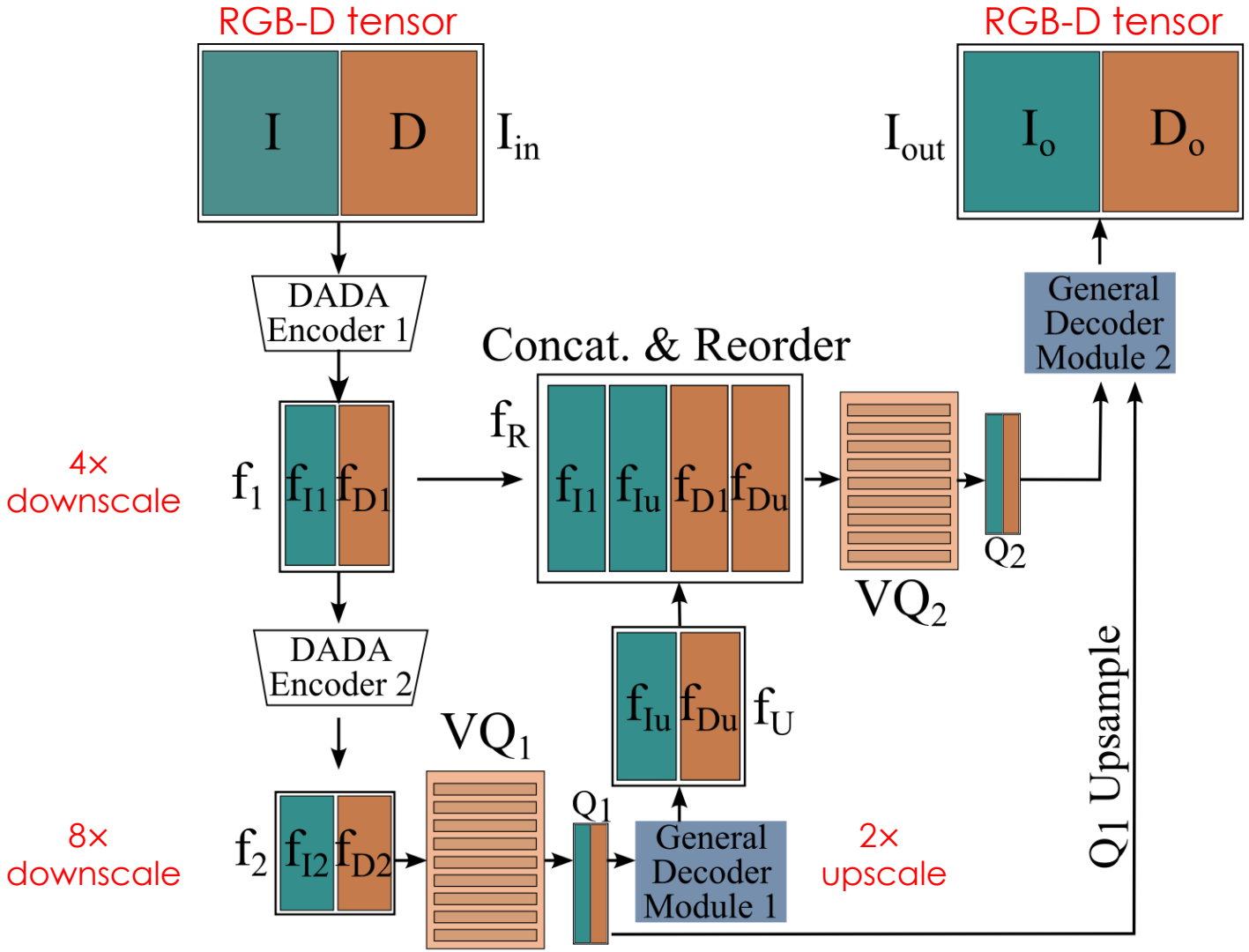
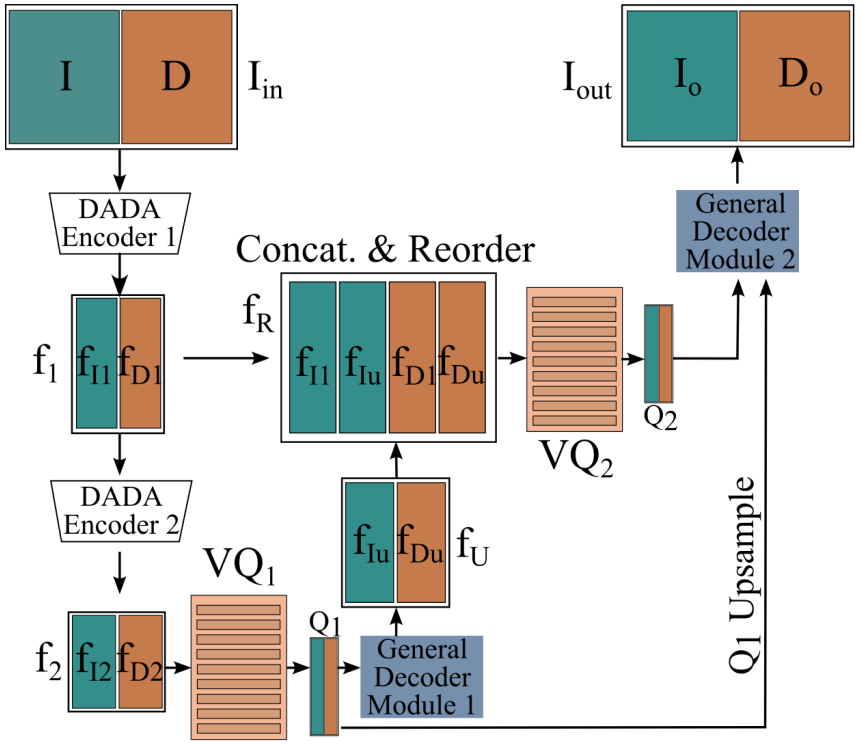


Figure 3. The Depth-Aware Discrete Autoencoder (DADA) module.

Training stage1 - DADA



$$\mathcal{L}_{recon} = \lambda_D L_2(D, D_o) + \lambda_I L_2(I, I_o)$$

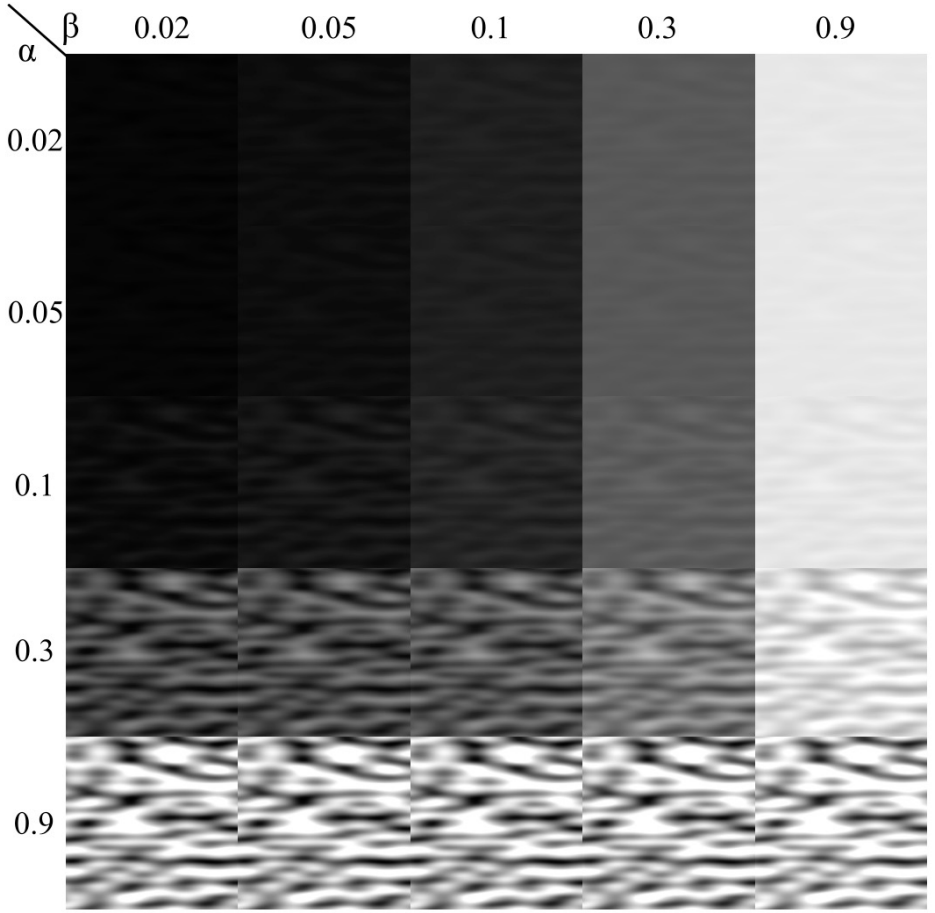
$$\mathcal{L}_{vq} = L_2(sg[f_2], Q_1) + L_2(sg[f_1], Q_2)$$

$$\mathcal{L}_{enc} = L_2(f_2, sg[Q_1]) + L_2(f_1, sg[Q_2])$$

$$\mathcal{L}_{ae} = \mathcal{L}_{recon} + \mathcal{L}_{vq} + \lambda_K \mathcal{L}_{enc} \qquad (1)$$

- Train DADA with RGB-D images for codebook generation
- RGB image: sampled from the ImageNet dataset
- Depth image: simulated from Perlin noise (unrelated with ImageNet)

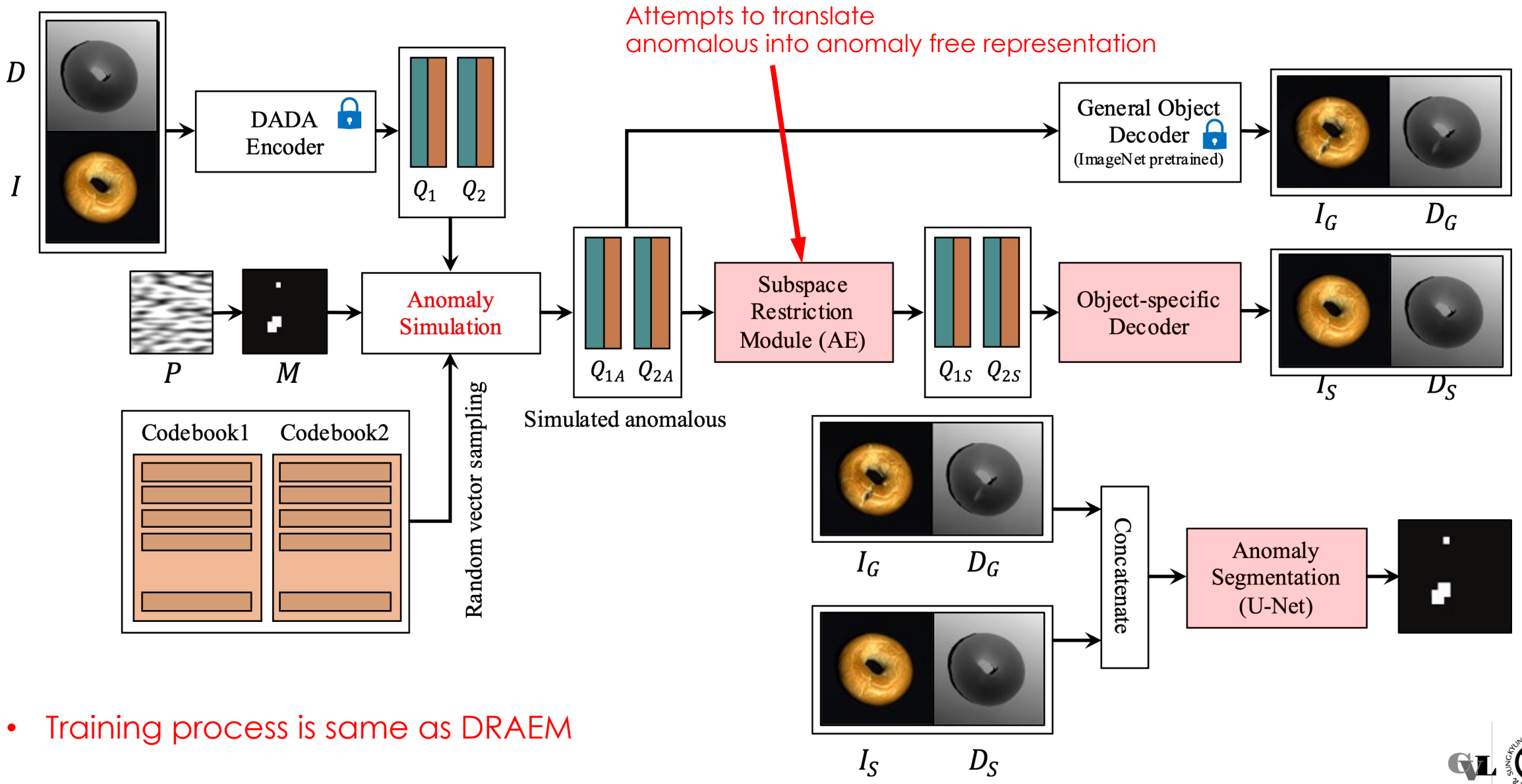
Inputs for DADA training



$$D = \alpha P + \beta$$

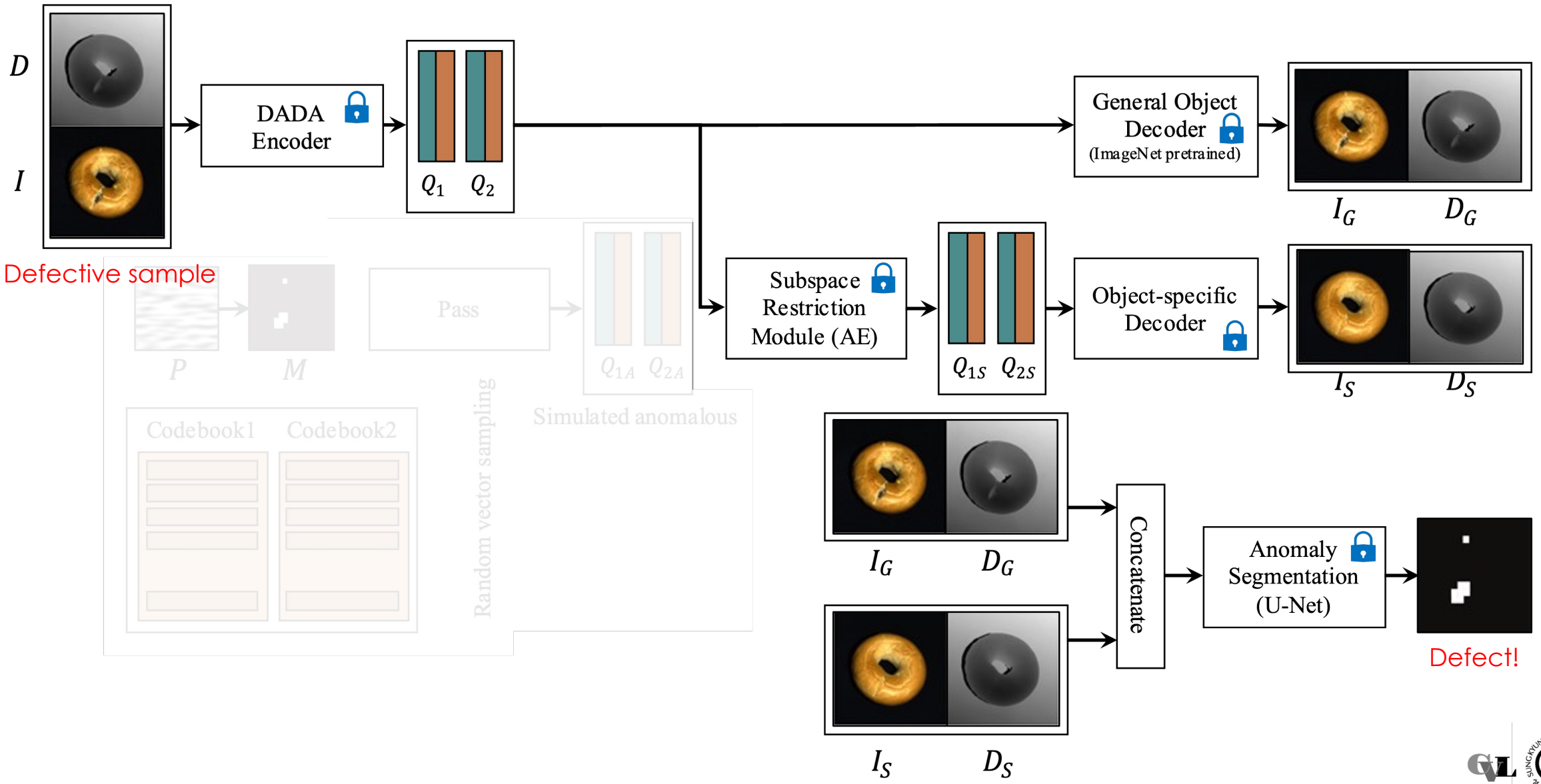
Figure 4. Depth image simulation with translation parameter α and β.

Training Stage2 - 3DSR



- Training process is same as DRAEM

Inference Stage



Paul Bergmann et al. "The MVTec 3D-AD dataset for unsupervised 3d anomaly detection and localization." arXiv. 2021.
Luca Bonfiglioli et al. "The eyecandies dataset for unsupervised multimodal anomaly detection and localization." ACCV. 202.

Qualitative Results

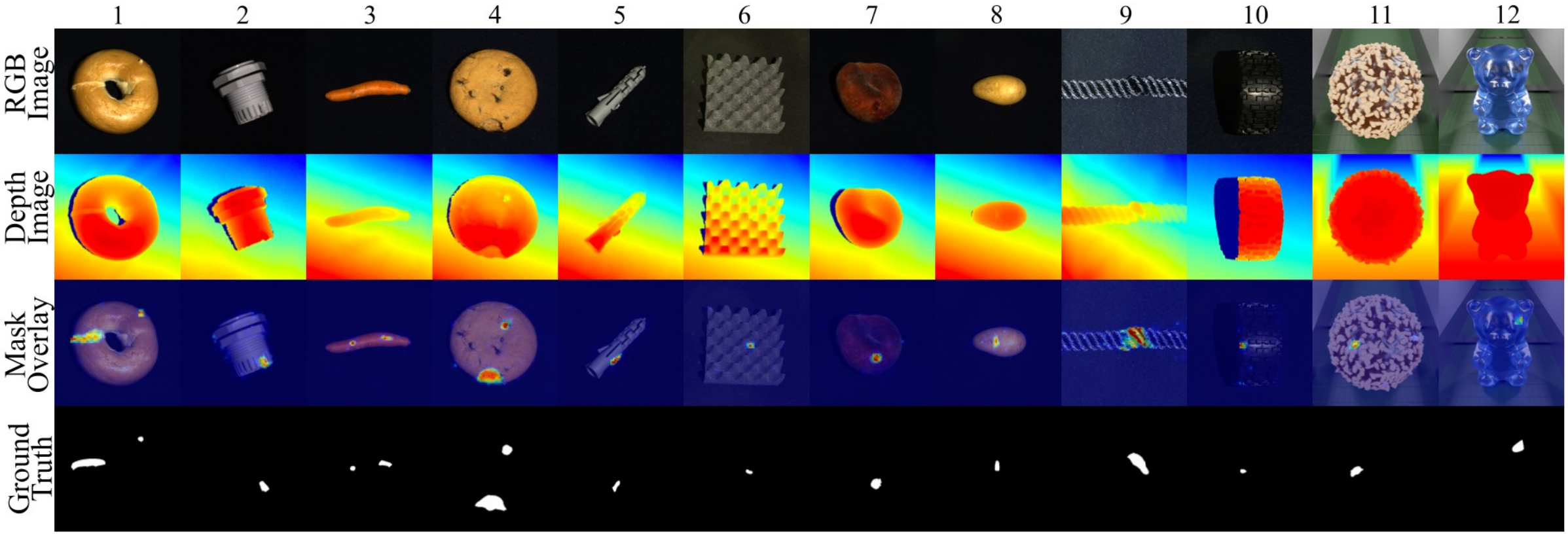
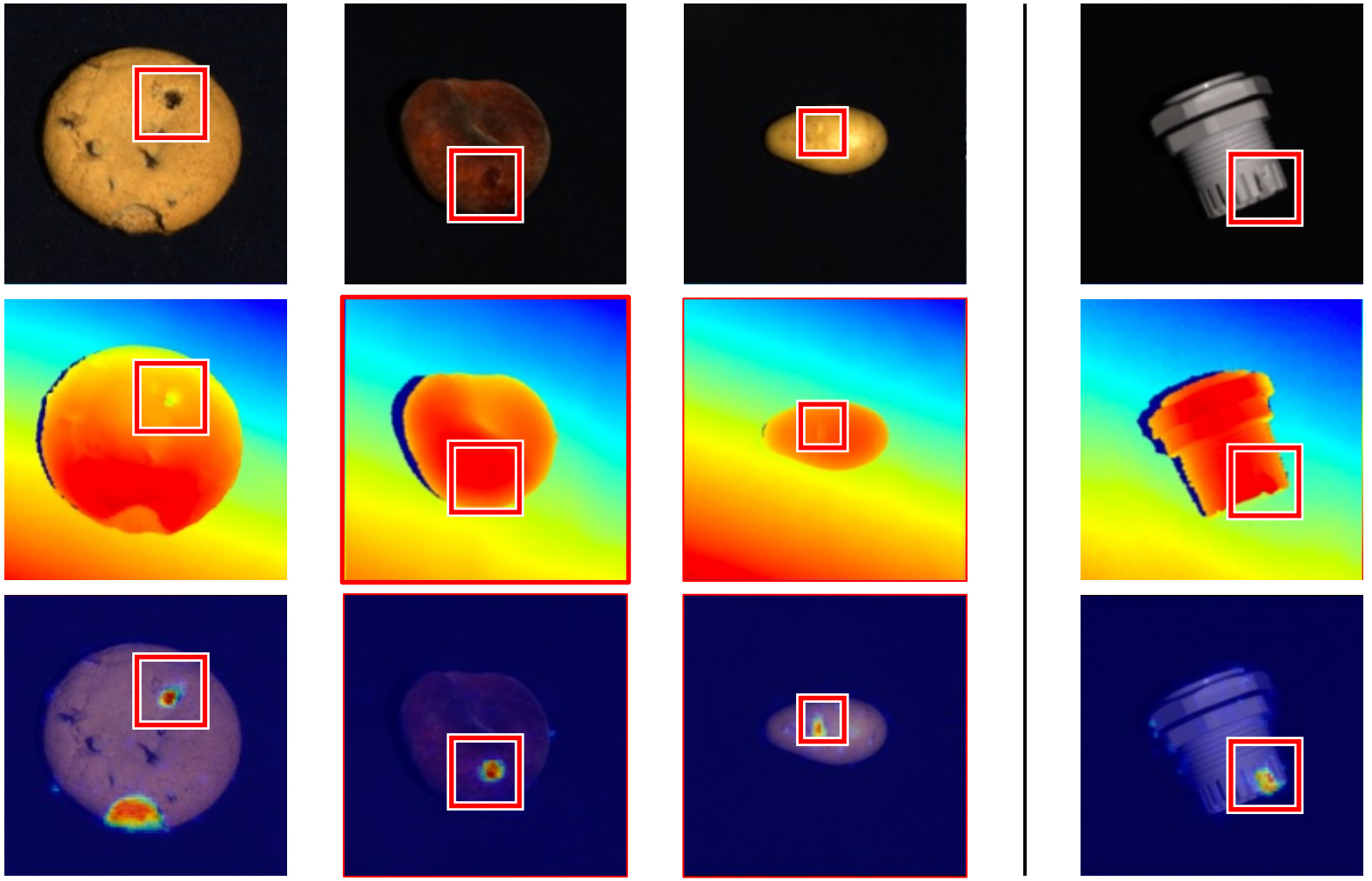


Figure 6. Qualitative results of 3DSR on the MVTec3D and Eyecandies benchmarks.

Complementary cases



Depth complementing RGB

RGB complementing Depth

Quantitative Results (1)

Table 1. Anomaly detection results on the MVTec3D dataset for the 3D, RGB and 3D+RGB problem setups. The results are listed as image-level AUROC scores (higher is better). The results of evaluated methods are ranked and the first, second and third place are marked.

	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
3D	Voxel AE [2]	69.3	42.5	51.5	79.0	49.4	55.8	53.7	48.4	63.9	58.3	57.1
	Depth GAN [2]	53.0	37.6	60.7	60.3	49.7	48.4	59.5	48.9	53.6	52.1	52.3
	Depth AE [2]	46.8	73.1②	49.7	67.3	53.4	41.7	48.5	54.9	56.4	54.6	54.6
	FPFH [9]	82.5	55.1	95.2	79.7	88.3③	58.2	75.8	88.9	92.9	65.3③	78.2
	3D-ST [3]	86.2	48.4	83.2	89.4③	84.8	66.3	76.3	68.7	95.8②	48.6	74.8
	AST_{3D} [17]	88.1③	57.6	96.5②	95.7②	67.9	79.7②	99.0①	91.5③	95.6③	61.1	83.3③
	$M3DM_{3D}$ [19]	94.1②	65.1③	96.5②	96.9①	90.5②	76.0③	88.0③	97.4①	92.6	76.5②	87.4②
	$3DSR_{3D}$	94.5①	83.5①	96.9①	85.7	95.5①	88.0①	96.3②	93.4③	99.8①	88.8 ①	92.2①
RGB	PatchCore [14]	87.6	88.0	79.1	68.2	91.2	70.1	69.5	61.8	84.1	70.2	77.0
	DifferNet [15]	85.9	70.3	64.3	43.5	79.7	79.0	78.7	64.3③	71.5	59.0	69.6
	PADiM [5]	97.5①	77.5	69.8	58.2	95.9	66.3	85.8	53.5	83.2	76.0	76.4
	CS-Flow [16]	94.1	93.0①	82.7	79.5②	99.0②	88.6③	73.1	47.1	98.6②	74.5	83.0
	AST_{RGB} [17]	94.7②	92.8③	85.1③	82.5①	98.1③	95.1①	89.5③	61.3	99.2①	82.1②	88.0②
	$M3DM_{RGB}$ [19]	94.4③	91.8	89.6②	74.9	95.9	76.7	91.9②	64.8②	93.8	76.7③	85.0③
	DSR_{RGB} [22]	84.4	93.0①	96.4①	79.4③	99.8①	90.4②	93.8①	73.0①	97.8③	90.0①	89.8①
3D+RGB	Voxel AE [2]	51.0	54.0	38.4	69.3	44.6	63.2	55.0	49.4	72.1	41.3	53.8
	Depth GAN [2]	53.8	37.2	58.0	60.3	43.0	53.4	64.2	60.1	44.3	57.7	53.2
	Depth AE [2]	64.8	50.2	65.0	48.8	80.5	52.2	71.2	52.9	54.0	55.2	59.5
	PatchCore+FPFH [9]	91.8	74.8	96.7	88.3	93.2	58.2	89.6	91.2③	92.1	88.6③	86.5
	AST [17]	98.3②	87.3②	97.6②	97.1③	93.2③	88.5③	97.4②	98.1①	100①	79.7	93.7③
	M3DM [19]	99.4①	90.9①	97.2③	97.6②	96.0②	94.2②	97.3③	89.9	97.2③	85.0③	94.5②
	3DSR	98.1③	86.7③	99.6①	98.1①	100①	99.4①	98.6①	97.8②	100①	99.5①	97.8①

Quantitative Results (2)

Table 4. Comparison between M3DM [19] and 3DSR on the Eyecandies dataset in terms of image-level AUROC (higher is better).

Method	Candy cane	Chocolate cookie	Chocolate praline	Confetto	Gummy Bear	Hazelnut truffle	Licorice sandwich	Lollipop	Marshmallow	Peppermint candy	Mean
$3DSR_{3D}$	60.0	76.8	74.2	77.0	76.1	74.9	81.1	83.1	81.1	91.7	**77.6**
$M3DM_{3D}$	48.2	58.9	80.5	84.5	78.0	53.8	76.6	82.7	80.0	82.2	72.5
DSR_{RGB} [22]	70.6	96.5	95.0	96.6	87.0	79.0	88.5	85.7	99.8	99.2	**89.8**
$M3DM_{RGB}$	64.8	94.9	94.1	100	87.8	63.2	93.3	81.1	998	100	87.9
$3DSR_{3D+RGB}$	65.1	99.8	90.4	97.8	87.5	86.1	96.5	89.9	99.0	97.1	**90.9**
$M3DM_{3D+RGB}$	62.4	95.8	95.8	100	88.6	75.8	94.9	83.6	100	100	89.7

Table 6. Method performance in terms of frames-per-second (FPS) on the NVIDIA RTX A4500 GPU.

Method	AST [17]	M3DM [19]	3DSR
FPS	18②	0.6③	33①

Ablation Study

Table 5. Ablation study results.

Method	I-AUROC	P-AUROC	PRO
DSR_{naive}	87.6	96.5	92.3
3DSR_{no_perlin}	90.0	98.3	93.3
3DSR_{no_affine}	94.8	99.2	95.9
3DSR_{VQVAE}	95.8	99.3	96.3
$\text{3DSR}_{weighted}$	96.5	99.4	96.7
3DSR	**97.8**	**99.5**	**97.2**

- DSR_{naive}: Training DADA with MVTec3D-AD
- 3DSR_{no_perlin}: Training DADA with RGB image and unpaired grayscale image from ImageNet
- 3DSR_{no_affine}: Training DADA without Perlin noise transformation
- 3DSR_{VQVAE}: Deactivating grouped convolution in DADA
  - Prevents the overwhelming influence of a single modality among RGB and depth
- $\text{3DSR}_{weighted}$: Removing weighting coefficient for training DADA
  - Just hyperparameter
- 3DSR: Full model

Summaries

Motivation and solution

- Motivation: Some surface anomalies are practically invisible in RGB space
- Solution: Integration of 3D information with RGB image

Contributions

- Achieve a high performance with small scale dataset
  - A fixed number of discrete latent representations of VQVAE enables training with a small-scaled dataset
  - Simulation process for learning informative depth-feature (training with synthetic data)
- 3DSR* to exploit joint modality information of RGB and depth
  - 3DSR Includes one DADA encoder, two latent decoders
  - DADA** learns joint representations of RGB and 3D data
  - State-of-the-art anomaly detection performance and processing time (w/ RTX A4500)

Remaining issues

- What if we/who cannot afford expensive 3D sensors?
- Why unrelated ImageNet-random simulated depth pair training is effective?

* 3DSR: 3D Dual Subspace Reprojection
** DADA: Depth-Aware Discrete Autoencoder

Appendix-A

TransFusion – A Transparency-Based Diffusion Model for Anomaly Detection

Matic Fučka, Vitjan Zavrtanik, Danijel Skočaj
University of Ljubljana, Faculty of Computer and Information Science
{matic.fucka, vitjan.zavrtanik, danijel.skocaj}@fri.uni-lj.si

Abstract

Surface anomaly detection is a vital component in manufacturing inspection. Reconstructive anomaly detection methods restore the normal appearance of an object, ideally modifying only the anomalous regions. Due to the limitations of commonly used reconstruction architectures, the produced reconstructions are often poor and either still contain anomalies or lack details in anomaly-free regions. Recent reconstructive methods adopt diffusion models, however with the standard diffusion process the problems are not adequately addressed. We propose a novel transparency-based diffusion process, where the transparency of anomalous regions is progressively increased, restoring their normal appearance accurately and maintaining the appearance of anomaly-free regions without loss of detail. We propose TRANSparency DifFUSION (TransFusion), a discriminative anomaly detection method that implements the proposed diffusion process, enabling accurate downstream anomaly detection. TransFusion achieves state-of-the-art performance on both the VisA and the MVTec AD datasets, with an image-level AUROC of 98.5% and 99.2%, respectively.

1. Introduction

The primary objective of surface anomaly detection is the identification and localization of anomalies in images. In the standard problem setup only anomaly-free (normal) images are used to learn a normal appearance model and any deviations from the learned model are classified as anomalies. Surface anomaly detection is commonly used in various industrial domains [8, 9, 45] where the limited availability of abnormal images along with their considerable diversity makes training supervised models impractical.

Many of the recent surface anomaly detection methods follow the reconstructive [1, 30, 37, 41] or the discriminative [40, 42] paradigms. Reconstructive methods train an autoencoder-like network on anomaly-free images and assume that the autoencoder will not generalize well to anomalous regions, since they were not seen during training, making them distinguishable by reconstruction error. Discriminative methods are trained to segment synthetic anomalies [38, 40, 44] and learn a normal-appearance model to generalize to real-world cases. A reconstructive network is commonly used as the normal-appearance model in discriminative methods.

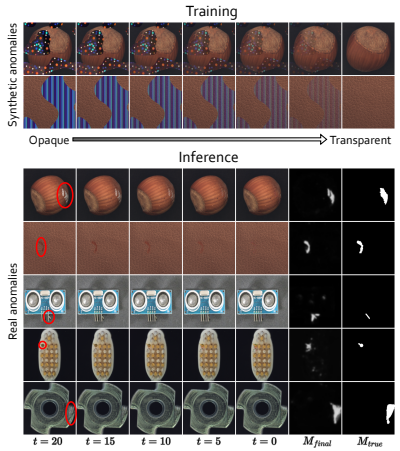


Figure 1. The reformulated diffusion model iteratively erases the anomalous regions during the backwards diffusion process. Training on synthetic anomalies (top) generalizes well to real anomalies (marked with red circles) seen at inference (bottom), leading to accurate output masks M_{final} that closely match the ground truth M_{true}.

Discriminative and reconstructive methods exhibit two core issues. First, reconstructive methods may *overgeneralize* which causes them to reconstruct even anomalous regions leading to false negative detections. Second, due to the limited image generation capabilities of the commonly

Overview

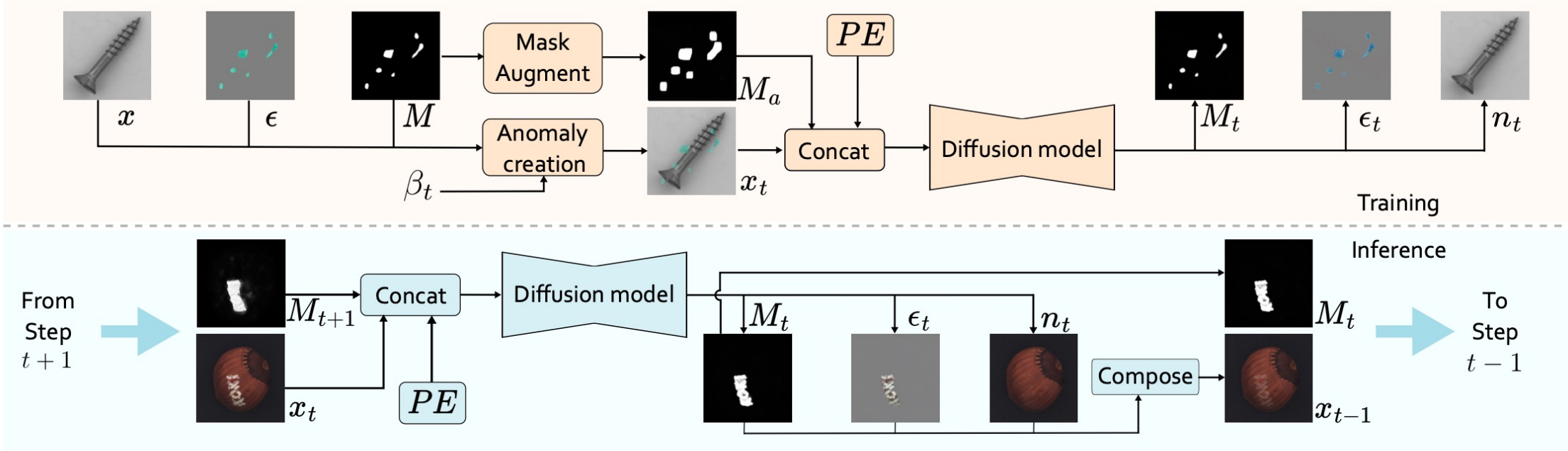


Figure 2. TransFusion's training and inference pipelines. Training examples are created from normal images x by generating the anomaly mask M and the anomaly appearance ϵ and imposing them on x according to the transparency schedule β_t. The resulting image x_t contains synthetic anomalies. TransFusion is guided by an augmented mask M_a. TransFusion outputs the estimated anomaly mask M_t, the anomaly appearance ϵ_t, and the normal appearance n_t. At inference, TransFusion infers M_t, ϵ_t, and n_t from the input image and constructs the next step image according to Eq. 4. The predicted mask M_t and the constructed x_{t-1} are used as the input in the next step.

Inference

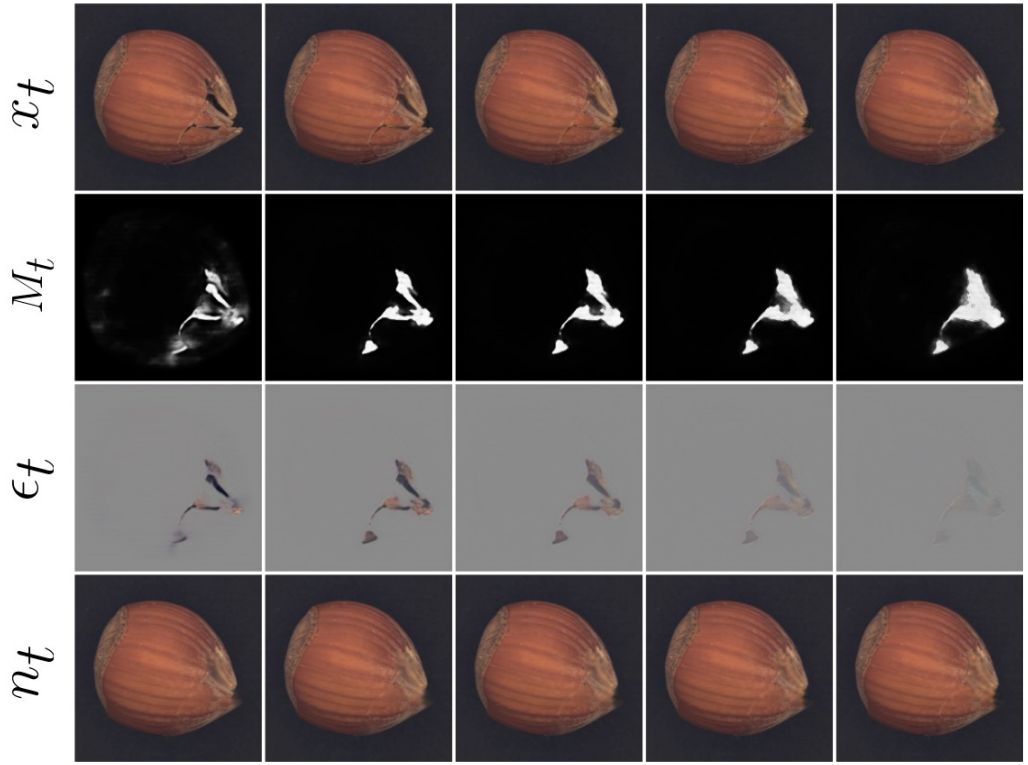


Figure 3. TransFusion inference. For every fourth timestep, the input image x_t and the predictions for the mask M_t, anomaly appearance ϵ_t and normal appearance n_t are shown. As seen in the top row TransFusion first reconstructs larger anomalies and inpaints the details near the end of the reconstruction process.

Method	DRÆM [40]	Patchcore [25]	DiffAD [43]	*TransFusion*
Inference [s]	**0.05**	0.22	1.00	0.34

Table 5. Results for average inference time of a single sample with NVIDIA A100 GPU. Inference times are reported in seconds.

Inference efficiency. Inference times of various methods can be seen in Table 5. Due to the complexity of diffusion models TransFusion is slower than some competing methods, however it is faster than other diffusion-based methods. Additionally, reducing the number of inference steps does not drastically reduce performance (Table 4). Diffusion distillation is an active field [23, 29, 32] and may be helpful for speeding up diffusion-based anomaly detection models.

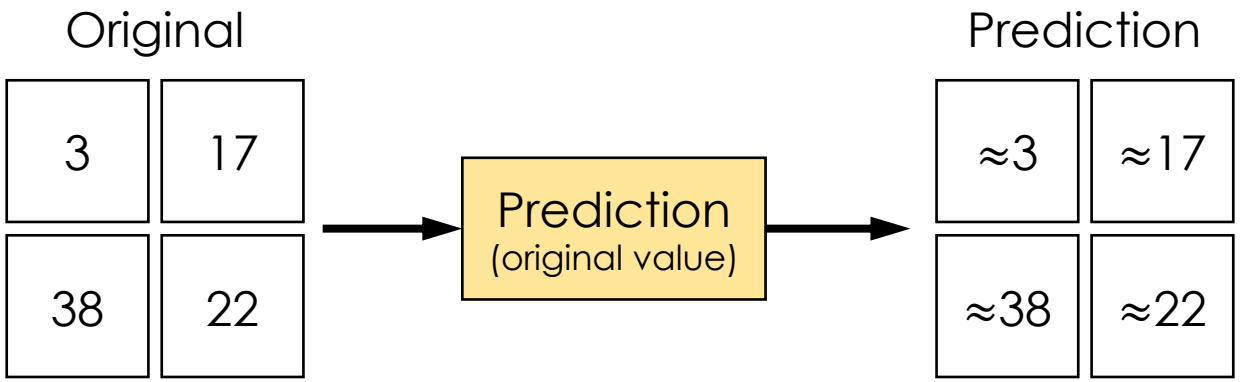
Performance

Method	AnoDDPM [37]	AnomDiff [22]	DiffAD [43]	DRÆM [40]	DSR [42]	FastFlow [39]	PatchCore [25]	RD4AD [11]	AST [27]	SimpleNet[21]	*TransFusion*
Candle	64.9	81.0	90.4	94.4	98.8	96.4	98.1	92.2	**99.4**	95.6	98.3
Capsules	76.5	80.0	87.6	76.3	99.1	89.2	85.7	90.1	85.4	76.7	**99.6**
Cashew	94.4	90.9	81.4	90.7	97.6	95.2	98.5	**99.6**	95.1	91.7	93.7
Chewing gum	91.3	98.1	94.0	94.2	93.8	99.4	99.0	99.7	**100**	99.1	99.6
Fryum	81.5	89.2	87.1	97.4	82.9	98.8	97.2	96.6	**99.1**	95.3	98.3
Macaroni1	58.8	77.8	87.6	95.0	87.3	94.5	95.7	**98.4**	93.9	90.8	**98.4**
Macaroni2	74.5	61.0	90.7	96.2	83.4	81.7	78.1	**97.6**	72.1	65.2	96.5
PCB1	42.1	86.7	75.0	54.8	90.5	94.7	98.3	97.6	**99.2**	60.1	98.9
PCB2	90.7	76.5	94.6	77.8	96.6	96.0	97.2	91.1	98.4	93.3	**99.7**
PCB3	92.3	80.4	94.7	94.5	94.8	93.3	96.2	95.5	97.4	94.9	**99.2**
PCB4	98.3	93.8	97.7	93.4	93.5	97.8	99.0	96.5	**99.6**	98.2	**99.6**
Pipe fryum	72.5	89.4	92.7	99.4	97.5	99.2	99.4	97.0	99.4	93.3	**99.6**
Average	78.2	83.7	89.5	88.7	91.6	93.9	94.3	96.0	94.9	87.9	**98.5**

Table 1. Comparison of TransFusion in anomaly detection (AUROC) with SOTA on VisA. First, second and third place are marked.

Appendix-B

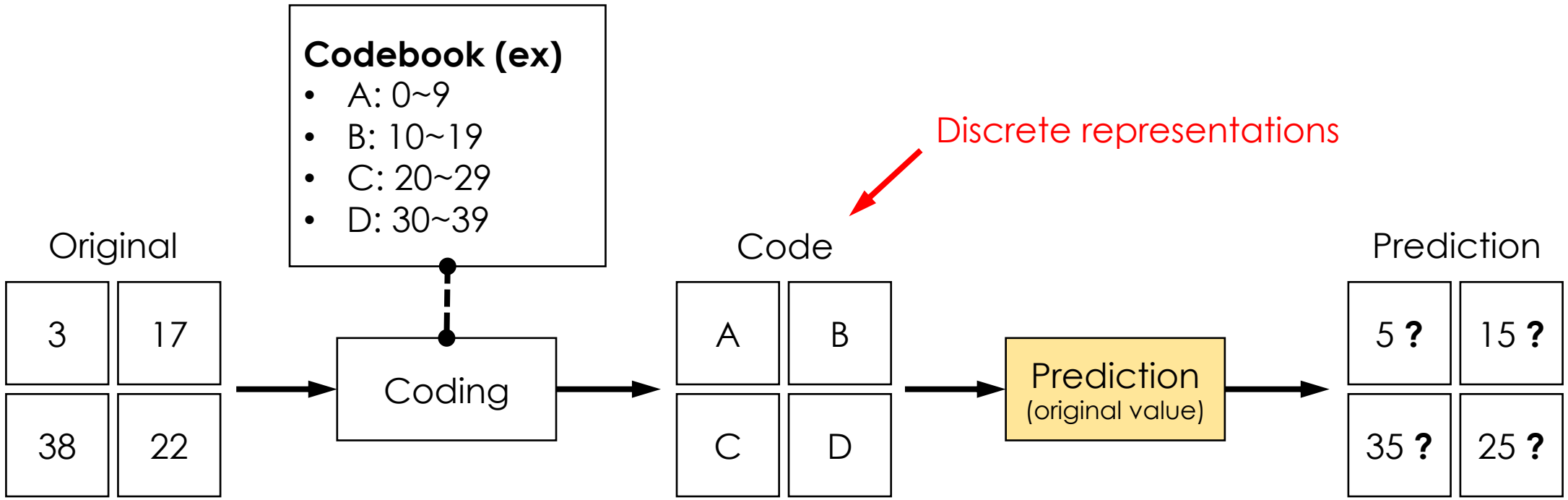
Why VQ?



- Similar to lossless compression & decompression
  - 1-to-1 matching between input and latent (continuous latent)
  - 1-to-1 matching of latent and prediction

- Easy to predict
  - No need to learn relationships between neighboring pixels

Why VQ?



- Similar to lossy compression & decompression
  - N-to-1 matching between input and latent (discrete latent)
  - 1-to-N matching of latent and prediction

- Not easy to predict
  - Need to learn relationships between neighboring pixels (as a hint for prediction)
  - Trained model will have context-aware strong representations