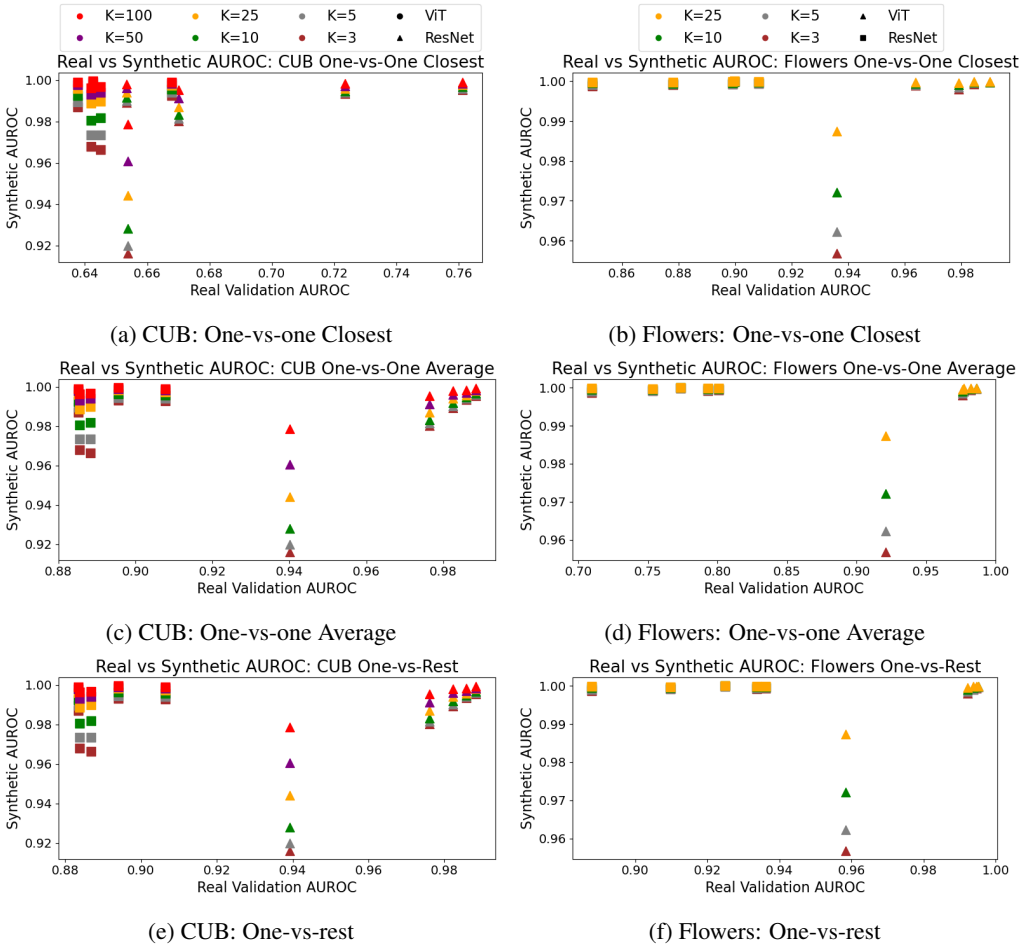


(a) CUB: One-vs-one Closest

(b) Flowers: One-vs-one Closest

(c) CUB: One-vs-one Average

(d) Flowers: One-vs-one Average

(e) CUB: One-vs-rest

(f) Flowers: One-vs-rest

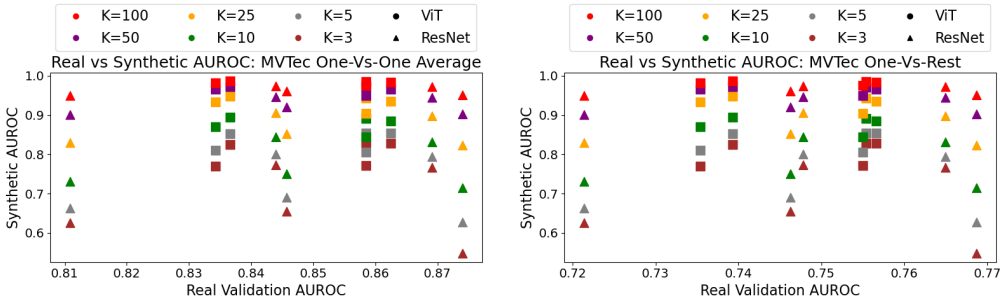
Figure 4: We compare the real validation AUROC against the AUROC when Tiny-Imagenet is used as synthetic anomalies, for the CUB (left) and Flowers (right) datasets and for all three benchmarks. We find that the performance is poor: model selection cannot be reliably performed when Tiny-ImageNet examples as used as synthetic anomalies.

A Appendix

A.1 Additional Results

Model selection with Tiny-ImageNet. Our initial experiments investigated if Tiny-ImageNet could be used effectively as synthetic anomalies when constructing the synthetic validation set. For these experiments, the same support set for each anomaly detection task is the same as the support for the corresponding experiment with our generated anomalies. When sampling anomalies from Tiny-Imagenet, we sample uniformly at random to generate a dataset $\tilde{X}_{\text{out}}$ of the same size: 100 images for tasks with the CUB and MVTec-AD datasets, and 25 images for tasks with the Flowers dataset. Ultimately, we found that using Tiny-ImageNet examples were not effective for our chosen tasks; in addition to the results for CLIP prompt selection in Table 1, the results for model selection are shown in Figure 4.

Additional MVTec-AD results. Figure 5 shows the results of the model selection experiment with the MVTec-AD dataset on the one-vs-one average and the one-vs-rest benchmarks (which were shown to be easier benchmarks to estimate in Figure 3). Unlike the CUB and Flowers datasets, in which synthetic anomalies could successfully approximate real validation performance, our synthetic anomalies are less effective for MVTec-AD.



(a) One-vs-One Average

(b) One-vs-Rest

Figure 5: Comparing the real validation AUROC against the synthetic validation AUROC for the MVTec-AD dataset. For all the one-vs-one average and the one-vs-rest benchmark, using our synthetic validation dataset is ineffective for selecting the true, best anomaly detection models.

We also comment on the performance of our method when selecting CLIP prompts for MVTec-AD. We noticed a disparity in performance between objects in MVTec-AD (e.g., capsules, cables, or screws) and textures in MVTec-AD (e.g., carpets, wood, or tiles). Our method is unable to select the best CLIP prompt in any of the six textures for MVTec-AD, instead only performing well on the nine MVTec-AD objects. We therefore identify yet another challenge when using our approach for MVTec-AD— DiffStyle relies on assumptions of "style" and "content" in their source images, and these elements are not present in MVTec-AD textures like carpets or tiles.

A.2 Generating Synthetic Outliers with Text-Guidance

A directional CLIP loss is defined using CLIP's image encoder E_I, CLIP's text encoder E_T, and source-target image and text pairs (x_{source}, x_{target}) and (y_{source}, y_{target}) respectively, StyleGAN-NADA enables text-guided generation of images:

$$\Delta T = E_T(y_{target}) - E_T(y_{source})$$
$$\Delta I = E_I(x_{target}) - E_I(x_{source})$$
$$L_{dir} = 1 - \frac{\Delta I \cdot \Delta T}{||\Delta I|| ||\Delta T||}$$

We use the Asyrp process (Kwon et al., 2023) for text-guided anomaly generation, but modify Asyrp in two ways: (i) using non-domain-specific text-guidance and (ii) defining the edit-strength of each anomaly. The original Asyrp process is evaluated on well-defined domains, and assumes that the source and target text are known (e.g, modifying "face" to "smiling face"). We instead propose a methox that does not assume a specific domain and does not require domain-specific texts as input.

First, we find that using a source text is unneeded, and a meaningful direction for ΔT can be extracted by using the image encoder E_I and the source image x_{source}. Second, we find that target texts can be replaced with auxiliary, out-of-domain texts, which we call y_{aux}. We make these two changes to redefine our directional loss L'_{dir}.

$$\Delta T' = E_T(y_{aux}) - E_I(x_{source})$$
$$\Delta I = E_I(x_{target}) - E_I(x_{source})$$
$$L'_{dir} = 1 - \frac{\Delta I \cdot \Delta T'}{||\Delta I|| ||\Delta T'||}$$

We follow the training procedure for h-space extractor defined in Kwon et al. (2023) to extract Δh, the direction in h-space that creates the desired change. We then apply Asyrp during the reverse diffusion process, but use the linear property of h-space and define an edit strength α, which indicates how strongly we modify the image. Referring to the formulation defined in Sec. 3.2, we modify the reverse DDIM process on the latent vector $x_T^{(1)}$ by adding our h-space term:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{P}_t(\epsilon_t^\theta(x_T^{(1)}|h^{(1)} + \alpha\Delta h)) + \mathbf{D}_t(\epsilon_t^\theta(x_T^{(1)})). \quad (3)$$

A.3 USING SYNTHETIC OUTLIERS FOR OUTLIER EXPOSURE

In addition to using synthetic outliers for validation, synthetic outliers can also be used for improving the performance of anomaly detection models through outlier exposure. Our methodology closely follows that of Fort et al. (2021); Mirzaei et al. (2023): we use a pre-trained vision transformer model, fine-tune the vision transformer on a surrogate classification task, and use distances in the trained embedding space as an anomaly detection.

Mirzaei et al. (2023) use a surrogate classification task for fine-tuning—a binary classification layer is added to the vision transformer, and the model is trained on benign in-class examples and synthetic outlier examples. In addition to the surrogate classification task, we also propose a regression-based task. We generate a variety of synthetic outliers with text-guidance, using varying edit strength α. When fine-tuning anomaly-detection models, the surrogate task is a regression that predicts α.

After fine-tuning, we remove the prediction head of the vision transformer. The support set is then converted into the transformer's embedding space (i.e., the last layer before the prediction layer) and used as a feature bank for anomaly detection. At test time, the total Euclidean distance to the closest three examples in the feature bank is used as the anomaly score.

A.4 CLIP PROMPT TEMPLATES

For our experiments in Sec. 4.3, we evaluated across set of ten candidate prompt templates. Our evaluated prompts are general-purpose, and only the term "bird" or "flower" is added to the template for the CUB and Flowers dataset respectively. For each dataset, the candidate prompt templates are provided below:

```
% CLIP Templates for Flowers
['a photo of a {} flower', 'a photo of some flower'],
['a cropped photo of a {} flower', 'a cropped photo of some flower'],
['a dark photo of a {} flower', 'a dark photo of some flower'],
['a photo of a {} flower for inspection', 'a photo of some flower for inspection'],
['a photo of a {} flower for viewing', 'a photo of some flower for viewing'],
['a bright photo of a {} flower', 'a bright photo of some flower'],
['a close-up photo of a {} flower', 'a close-up photo of some flower'],
['a blurry photo of a {} flower', 'a blurry photo of some flower'],
['a photo of a small {} flower', 'a photo of a small some flower'],
['a photo of a large {} flower', 'a photo of a large some flower'],
```

```
% CLIP Templates for CUB
['a photo of a {} bird', 'a photo of some bird'],
['a cropped photo of a {} bird', 'a cropped photo of some bird'],
['a dark photo of a {} bird', 'a dark photo of some bird'],
['a photo of a {} bird for inspection', 'a photo of some bird for inspection'],
['a photo of a {} bird for viewing', 'a photo of some bird for viewing'],
['a bright photo of a {} bird', 'a bright photo of some bird'],
['a close-up photo of a {} bird', 'a close-up photo of some bird'],
['a blurry photo of a {} bird', 'a blurry photo of some bird'],
['a photo of a small {} bird', 'a photo of a small some bird'],
['a photo of a large {} bird', 'a photo of a large some bird'],
```

```
% CLIP Templates for MVTec
['a photo of a {}', 'a photo of something'],
['a cropped photo of a {}', 'a cropped photo of something'],
['a dark photo of a {}', 'a dark photo of something'],
['a photo of a {} for inspection', 'a photo of something for inspection'],
['a photo of a {} for viewing', 'a photo of something for viewing'],
['a bright photo of a {}', 'a bright photo of something'],
['a close-up photo of a {}', 'a close-up photo of something'],
['a blurry photo of a {}', 'a blurry photo of something'],
['a photo of a small {}', 'a photo of a small something'],
['a photo of a large {}', 'a photo of a large something'],
```