Table 3. Sub-Image anomaly detection accuracy on MVTec (PRO %)

	Student	1-NN	OC-SVM	ℓ_2-AE	VAE	SSIM-AE	CNN-Dict	SPADE
Carpet	69.5	51.2	35.5	45.6	50.1	64.7	46.9	94.7
Grid	81.9	22.8	12.5	58.2	22.4	84.9	18.3	86.7
Leather	81.9	44.6	30.6	81.9	63.5	56.1	64.1	97.2
Tile	91.2	82.2	72.2	89.7	87.0	17.5	79.7	75.9
Wood	72.5	50.2	33.6	72.7	62.8	60.5	62.1	87.4
Bottle	91.8	89.8	85.0	91.0	89.7	83.4	74.2	95.5
Cable	86.5	80.6	43.1	82.5	65.4	47.8	55.8	90.9
Capsule	91.6	63.1	55.4	86.2	52.6	86.0	30.6	93.7
Hazelnut	93.7	86.1	61.6	91.7	87.8	91.6	84.4	95.4
Metal nut	89.5	70.5	31.9	83.0	57.6	60.3	35.8	94.4
Pill	93.5	72.5	54.4	89.3	76.9	83.0	46.0	94.6
Screw	92.8	60.4	64.4	75.4	55.9	88.7	27.7	96.0
Toothbrush	86.3	67.5	53.8	82.2	69.3	78.4	15.1	93.5
Transistor	70.1	68.0	49.6	72.8	62.6	72.5	62.8	87.4
Zipper	93.3	51.2	35.5	83.9	54.9	66.5	70.3	92.6
Average	85.7	64	47.9	79	63.9	69.4	51.5	**91.7**

Table 4. Image-level anomaly detection accuracy on STC (Average ROCAUC %)

TSC [23]	StackRNN [23]	AE-Conv3D [35]	MemAE [12]	AE(2D) [16]	ITAE [19]	SPADE
67.9	68.0	69.7	71.2	60.9	**72.5**	71.9

In Tab. 3, we compare our method in terms of PRO. As explained above, this is another per-pixel accuracy measure which gives larger weight to anomalies which cover few pixels. Our method is compared with the auto-encoder with pre-trained features based approach of Bregmann et al. [6] and the baselines presented in their paper. Our approach achieves significantly better results than all previous methods. We note than Bregmann et al also presented an ensemble approach with better results. While our method does not use ensembles (which will probably improve our method too), we outperform the ensemble approach as well. We present more qualitative results of our method in Fig. 1 that show that our method is able to recover accurate masks of the anomalous regions.

4.2 Shanghai Tech Campus Dataset

We evaluate our method on the Shanghai Tech Campus dataset. It simulates a surveillance setting, where the input consists of videos captured by surveillance cameras observing a busy campus. The dataset contains 12 scenes, each scene consists of training videos and a smaller number of test images. The training videos do not contain anomalies while the test videos contain normal and anomalous images. Anomalies are defined as pedestrians performing non-standard ac-