# APPENDIX

## A. New Datasets

Besides the datasets from category “new datasets” in the section 4.1.2, we also use:

**CIFAR-10-DDPM-ema,** we sample 2,000 images using a pre-trained DDPM model[3]. As a real dataset, we employ the images from the CIFAR-10 dataset [46].

**Oxford-Flowers-64-DDPM-ema,** we sample 2,000 images using the pre-trained DDPM model from diffusers [90], with the id “flowers-102-categories”. As a real dataset, we employ the images from the diffuser dataset with the id “huggan/flowers-102-categories”.

**CelebaHQ-256-{DDPM, DDIM, PNDM, LDM}-ema,** we sample 2,000 images (for each metthod) using pre-trained DDPM, DDIM, PNDM and LDM models from diffusers [90], with the id “google/ddpm-ema-celebahq-256” and “CompVis/ldm-celebahq-256”, respectively. As real dataset, we employ the images from CelebaHQ dataset [42] from kaggle[4], which already provides the dimensions 256 × 256 pixels.

**LSUN-Cat-{DDPM, DDIM, PNDM}-ema,** we sample 2,000 images (for each method) using pre-trained DDPM, DDIM and PNDM models from diffusers [90], with the id “google/ddpm-ema-cat-256”. As a real dataset, we employ the images from the original source[5] [98]. We center-crop them to 256 × 256 pixels.

**LSUN-Church-{DDPM, DDIM, PNDM}-ema,** we sample 2,000 images (for each method) using pre-trained DDPM, DDIM and PNDM models from diffusers [90], with the id “google/ddpm-ema-church-256”. As a real dataset, we employ the images from the original source [98]. We center-crop them to 256 × 256 pixels.

**ImageNet-DiT,** we sample 2,000 images using a pre-trained DiT model from diffusers [90], with the id “facebook/DiT-XL-2-256” [61]. As a real dataset, we employ the images from the original source [19]. We center-crop them to 256 × 256 pixels.

## B. Definition of LID

This section extends the explanation of LID in section 3.1: Let $\mathbb{R}^m$ denote a continuous domain with a non-negative distance function $d$. The continuous intrinsic dimensionality aims to measure the local intrinsic dimensionality of $\mathbb{R}^m$ based on the distribution of interpoint distances. For a fixed point $x$, the distribution of distances can be represented as a random variable $\mathbf{D}$ on $[0, +\infty)$ with a probability density function $f_D$ and cumulative density function $F_D$.

When considering samples $x$ drawn from continuous probability distributions, the intrinsic dimensionality is defined as follows [3]:

**Definition 5.1** *Intrinsic Dimensionality (ID). Given a sample $x \in \mathbb{R}^m$, let $D$ be a random variable denoting the distance from $x$ to other data samples. If the cumulative distribution $F(d)$ of $\mathbf{D}$ is positive and continuously differentiable at distance $d > 0$, the ID of $x$ at distance $d$ is given by:*

$$\mathrm{ID}_{\mathbf{D}}(d) \triangleq lim_{\epsilon\to 0}\frac{\log F_{\mathbf{D}}((1+\epsilon)d) - \log F_{\mathbf{D}}(d)}{\log(1+\epsilon)} \tag{5}$$

In practice, we are given a fixed number $n$ of samples of $x$, allowing us to compute their distances to $x$ in ascending order $d_1 \leq d_2 \leq \cdots \leq d_{n-1}$, with a maximum distance between any two samples. As shown in [3], the log-likelihood of $\mathrm{ID}_{\mathbf{D}}(d)$ for $x$ is given as:

$$n\log\frac{F_{\mathbf{D},w}(w)}{w} + n\log\mathrm{ID}_{\mathbf{D}} + (\mathrm{ID}_{\mathbf{D}} - 1)\sum_{i=1}^{n-1}\log\frac{d_i}{w}. \tag{6}$$

The maximum likelihood estimate is then given by:

$$\widehat{\mathrm{ID}}_{\mathbf{D}} = -\left(\frac{1}{n}\sum_{i=0}^{n-1}\log\frac{d_i}{w}\right)^{-1} \qquad \text{with} \tag{7}$$

$$\widehat{\mathrm{ID}}_{\mathbf{D}} \sim \mathcal{N}\left(\mathrm{ID}_{\mathbf{D}}, \frac{\mathrm{ID}_{\mathbf{D}}^2}{n}\right), \tag{8}$$

meaning that the estimate is drawn from a normal distribution with a mean of $\mathrm{ID}_{\mathbf{D}}$ and a variance that decreases linearly with an increasing number of samples, while it increases quadratically with $\mathrm{ID}_{\mathbf{D}}$. The *local* ID is an estimation of the intrinsic dimension based on the local neighborhood of a point $x$, such as its $k$ nearest neighbors, as shown in equation (1).

## C. Un/trained Feature Maps

In table 3, we show the comparative analysis if we calculate the multiLID on untrained and trained feature maps. To unveil differences, we use logistic regression (LR) as a second binary classifier besides random forest (RF). On the dataset Cifar-10 the untrained feature maps exhibits slightly increasing detection accuracy. On the ImageNet dataset, which has a hierarchical class structure and larger resolutions is not that much difference to observe between trained and untrained feature maps.

[3] https://github.com/pesser/pytorch_diffusion

[4] https://www.kaggle.com/datasets/denislukovnikov/celebahq256-images-only

[5] https://www.yf.io/p/lsun

Table 3: This table shows an ablation study when extracting the features of an *untrained* and a *trained* ResNet18. To lift the insights, we evaluated besides the random forest (RF), also the logistic regression (LR) classifier.

| **dataset** | **gen. model** | **untrained** | | **trained** | |
|---|---|---|---|---|---|
| | | RF | LR | RF | LR |
| Cifar-10 | ddpm ema | 1.0 | 0.67 | 1.0 | 0.77 |
| ImageNet | dit-xl-2 | 1.0 | 0.71 | 1.0 | 0.69 |

## D. Variance of the Strength Assessment

In this section, we show additionally to the strength assessment of the multiLID (see fig. 4), the variance over 5 runs per tile (see fig. 7). This ablation study of multiLID shows the variance of the accuracy rates when using different numbers of samples and accumulating the features (from previous to later layers). The maximum variance is around $10^{-3}$ and becomes 0 when the number of samples is larger than 800 per class.

## E. Robustness via Data Augmentation

In this section, we extend the section 4.3 by evaluating Gaussian blurring and JPEG compression on more datasets.

We extend the data augmentation evaluation from fig. 3 in the section 4.3 by using more datasets: i.e. CelbeaHQ (fig. 8), LSUN-Cat (fig. 9), LSUN-Church (fig. 10), and LSUN-Bedroom (fig. 11).

Furthermore, we extend our experiments by using a standardized augmented training procedure by mixing the two-class degradation and using different parameters randomly. Similar to [93], our images are randomly Gaussian blurred with $\sigma \sim \text{Uniform}[0, 3]$ and compressed with a quality $\sim \text{Uniform}\{30, 31, \ldots, 100\}$. We conduct three independent experiments: i) No augmentation: Trained and tested on clean data. We report accuracy (ACC) as an evaluation metric. ii) Moderate augmentation: Images are randomly Gaussian blurred and compressed with the JPEG algorithm. The augmentation probability is set to 0.5. iii) Strong augmentation: Likewise previous augmentation, but with a probability greater than 0.1. We can observe in the table 4 that with data augmentation our approach based on multiLID is able to yield accurate detection results on all deterioration, i.e. Gaussian blur and JPEG compression.

## F. Limitation of Identification and Transferability

In this section, we extend the evaluation in section 4.5 by the datasets CelebaHQ (fig. 12), LSUN-Cat (fig. 13), and LSUN-Church (fig. 14). Analogous to LSUN-Bedroom (fig. 5), the other datasets also depict similar identification and transfer capabilities. Finally, we add to the identification of the Artifact dataset in the fig. 5 the transferability in fig. 15.

## G. Feature Importance

The feature importance[6] helps us in understanding which features have the most significant impact on the model's performance. More specifically, the importance is calculated based on how much each feature contributes to reducing the impurity or error of the model. In the context of random forest classifier [9], this method provides a feature importance score as a byproduct of its training process. In this case, each selected ResNet18 layer $\ell$ represents a feature. Note that the sum over all layers is 1, i.e. $\sum_{\ell=1}^{8} |f_\ell| = 1$. In our implementation, we use the Gini importance, also known as mean decrease in impurity (MDI) [55]. This method calculates each feature importance as the sum of the number of splits across all trees that include the feature, proportionally to the number of samples it splits. In fig. 6, we display the feature importance of each extracted ReLU layer from our ResNet18. We can confirm the observation from [30], that the first ReLU layer (the shallowest) is the least significant, while the last ReLU layer (the deepest) is the most important across all our benchmark datasets.

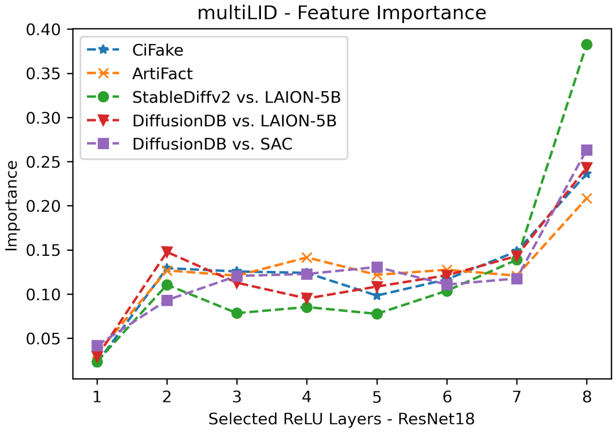


Figure 6: Feature importance from our classifier. The features are extracted per sample after each ReLU activation from an untrained ResNet18. As it can be noticed, the last layer plays a crucial role, in contrast to the first one.

[6] https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

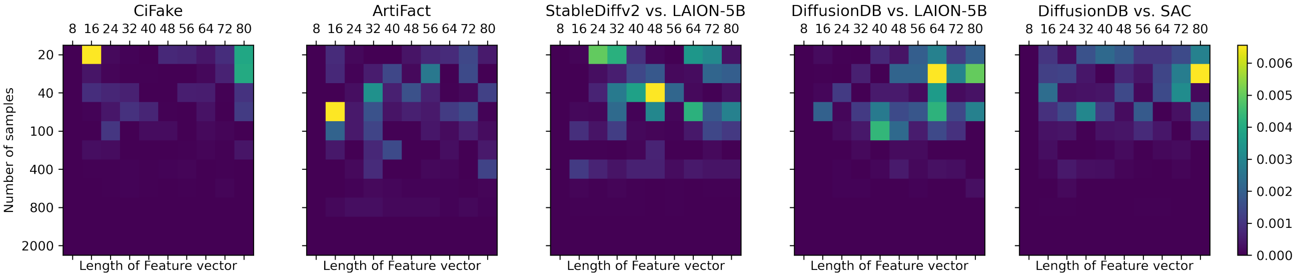


Figure 7: Ablation study of the variance (see section D) multiLID detection accuracy by using different numbers of samples and accumulating the features (from previous to later layers) and extending the strength evaluation in fig. 4. The variance reaches confidently zero by increasing the number of training samples.

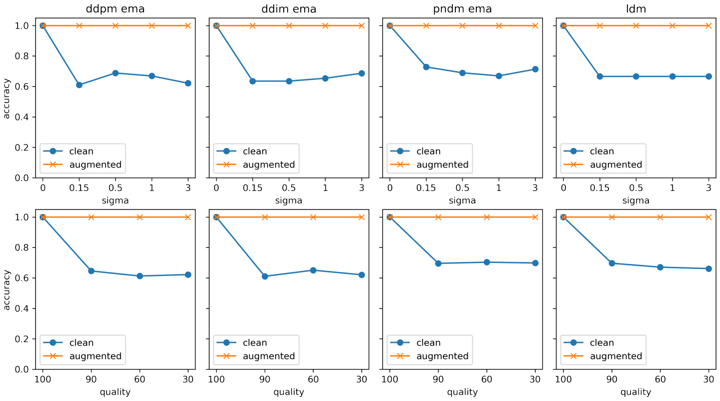


Figure 8: Data augmentation on the CelebaHQ models. Robustness (see section E) of Gaussian blurring (top row) and JPEG compression (bottom row). In both cases the data augmentation is necessary to improve the detectors' accuracy.

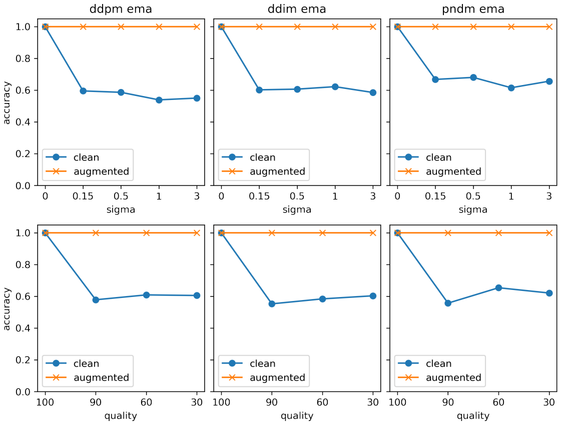


Figure 9: Robustness (see section E) against Gaussian blurring (top row) and JPEG compression (bottom row) on the LSUN-Cat datasets. In both cases the data augmentation is necessary to improve the detectors' accuracy.

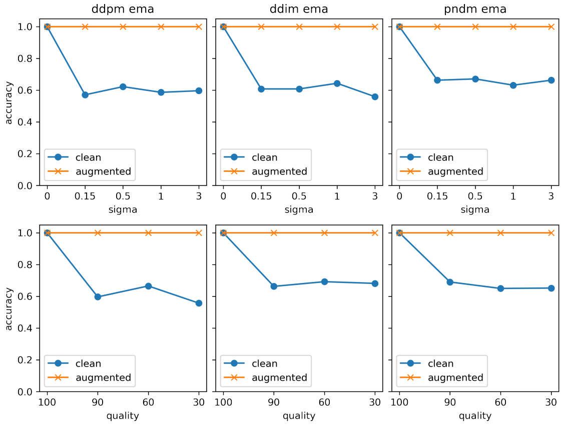


Figure 10: Robustness (see section E) against Gaussian blurring (top row) and JPEG compression (bottom row) on the LSUN-Church datasets. In both cases the data augmentation is necessary to improve the detectors' accuracy.

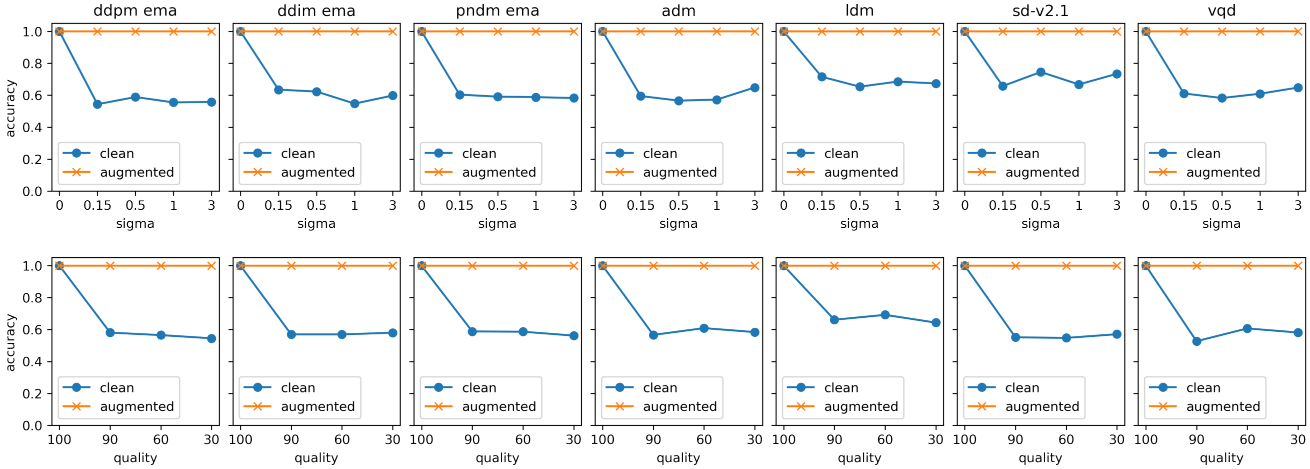
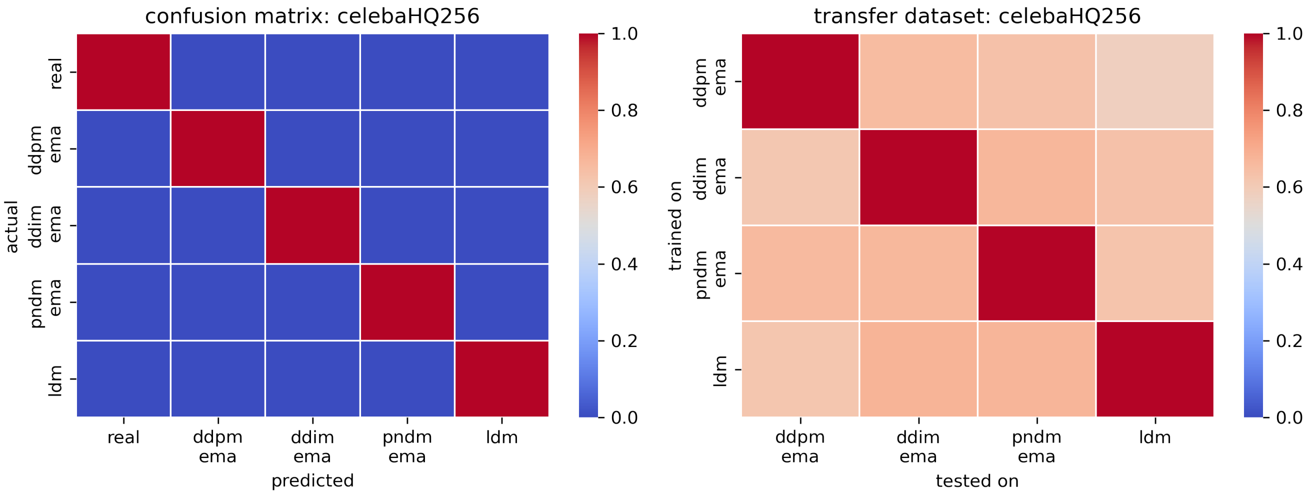


Figure 11: Robustness (see section E) against Gaussian blurring (top row) and JPEG compression (bottom row) on the LSUN-Bedroom datasets. In both cases the data augmentation is necessary to improve the detectors' accuracy.

Table 4: Data augmentation (Gaussian blurring and JPEG compression inspired from [93]) on different datasets. To evaluate the multiLID, we use as measurement the accuracy (ACC). While the classifier trained and evaluated on clean data shows accurate detection results, the accuracy drops by using Gaussian-blurred or JPEG-compressed data on the classifier trained on clean data. Further details in the section E.

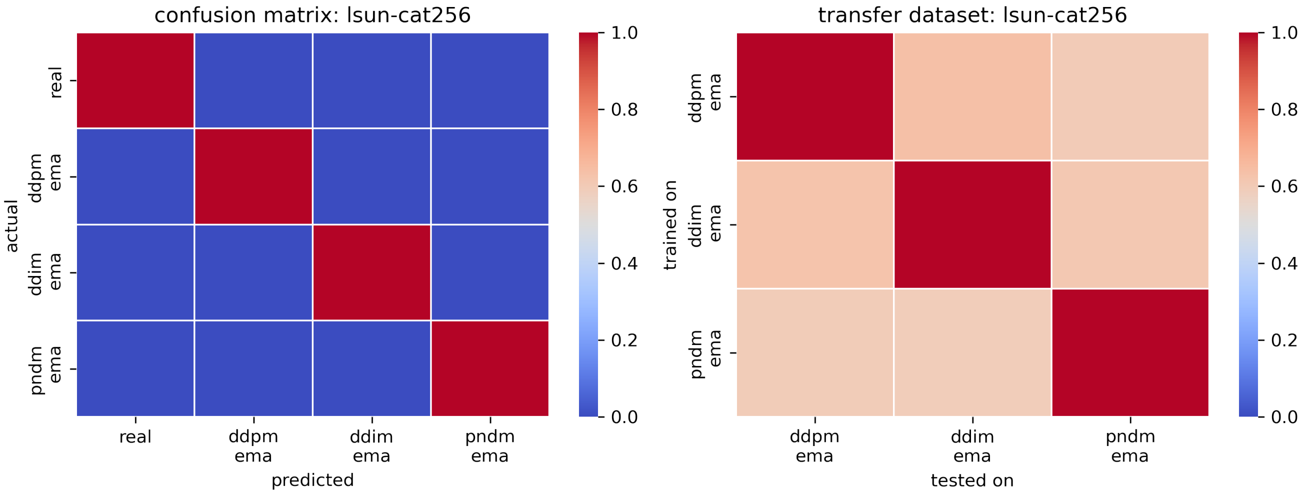
| **dataset** | **model** | **size** | **multiLID (ACC)** | | | | |
|---|---|---|---|---|---|---|---|
| | | | clean | blur+JPEG (0.5) | | blur+JPEG (0.1) | |
| | | | | clean | robust | clean | robust |
| CiFake | | 32 | 1.0 | 0.696 | 1.0 | 0.638 | 1.0 |
| ArtiFact | | 200 | 1.0 | 0.598 | 1.0 | 0.569 | 1.0 |
| SD-v2.1 vs. LAION-5B | | 768 | 1.0 | 0.714 | 1.0 | 0.641 | 1.0 |
| DiffusionDB vs. LAION-5B | | 512 | 1.0 | 0.644 | 1.0 | 0.657 | 1.0 |
| DiffusionDB vs. SAC | | 512 | 1.0 | 0.602 | 1.0 | 0.672 | 1.0 |
| Cifar-10 | DDPM ema | 32 | 1.0 | 0.602 | 1.0 | 0.567 | 1.0 |
| Oxford Flowers 102 | DDPM ema | 64 | 1.0 | 0.592 | 1.0 | 0.524 | 1.0 |
| CelebaHQ-256 | DDPM ema | 256 | 1.0 | 0.551 | 1.0 | 0.584 | 1.0 |
| | DDIM ema | 256 | 1.0 | 0.576 | 1.0 | 0.531 | 1.0 |
| | PNDM ema | 256 | 1.0 | 0.654 | 1.0 | 0.562 | 1.0 |
| | LDM | 256 | 1.0 | 0.644 | 1.0 | 0.594 | 1.0 |
| LSUN-Cat | DDPM ema | 256 | 1.0 | 0.651 | 1.0 | 0.602 | 1.0 |
| | DDIM ema | 256 | 1.0 | 0.586 | 1.0 | 0.510 | 1.0 |
| | PNDM ema | 256 | 1.0 | 0.580 | 1.0 | 0.600 | 1.0 |
| LSUN-Church | DDPM ema | 256 | 1.0 | 0.564 | 1.0 | 0.584 | 1.0 |
| | DDIM ema | 256 | 1.0 | 0.662 | 1.0 | 0.618 | 1.0 |
| | PNDM ema | 256 | 1.0 | 0.656 | 1.0 | 0.634 | 1.0 |
| LSUN-Bedroom | DDPM ema | 256 | 1.0 | 0.600 | 1.0 | 0.549 | 1.0 |
| | DDIM ema | 256 | 1.0 | 0.644 | 1.0 | 0.594 | 1.0 |
| | PNDM ema | 256 | 1.0 | 0.590 | 1.0 | 0.537 | 1.0 |
| | ADM | 256 | 1.0 | 0.584 | 1.0 | 0.600 | 1.0 |
| | LDM | 256 | 1.0 | 0.614 | 1.0 | 0.656 | 1.0 |
| | SD-v2.1 | 256 | 1.0 | 0.622 | 1.0 | 0.656 | 1.0 |
| | VQD | 256 | 1.0 | 0.576 | 1.0 | 0.542 | 1.0 |



(a) Identification.

(b) Transferability.

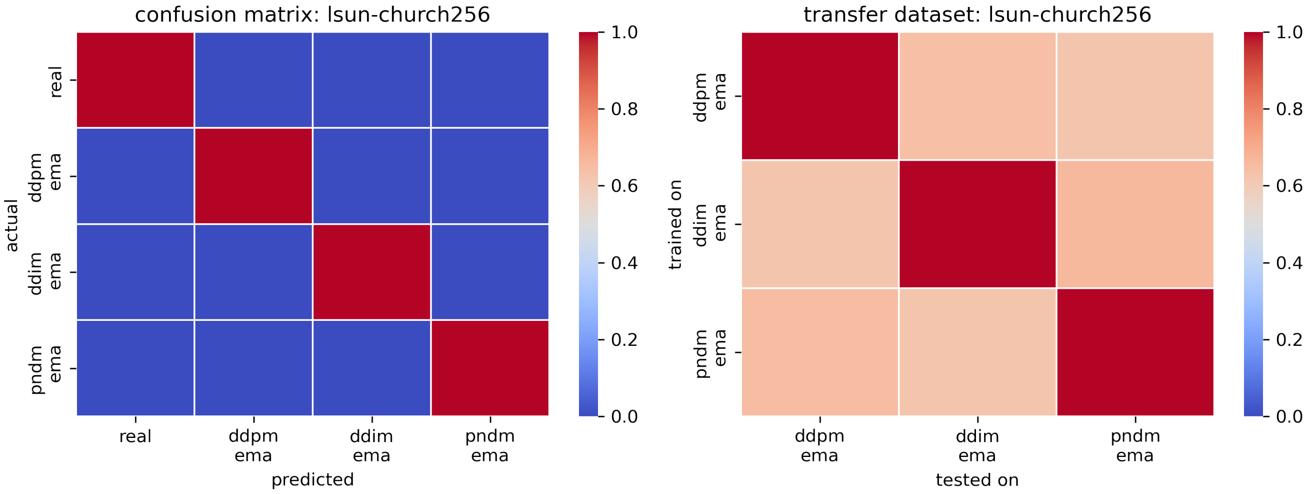
Figure 12: Identficiation and transferability on the CelebaHQ datasets described in section 4.1. Analogous to the experiments on the LSUN-Bedroom in section 4.5, the identification is accurate while the transferability is rather low.



(a) Identification.

(b) Transferability.

Figure 13: Identficiation and transferability on the LSUN-Cat datasets described in section 4.1. Analogous to the experiments on the LSUN-Bedroom in section 4.5, the identification is accurate while the transferability is rather low.



(a) Identification.

(b) Transferability.

Figure 14: Identficiation and transferability on the LSUN-Church datasets described in section 4.1. Analogous to the experiments on the LSUN-Bedroom in section 4.5, the identification is accurate while the transferability is rater low.

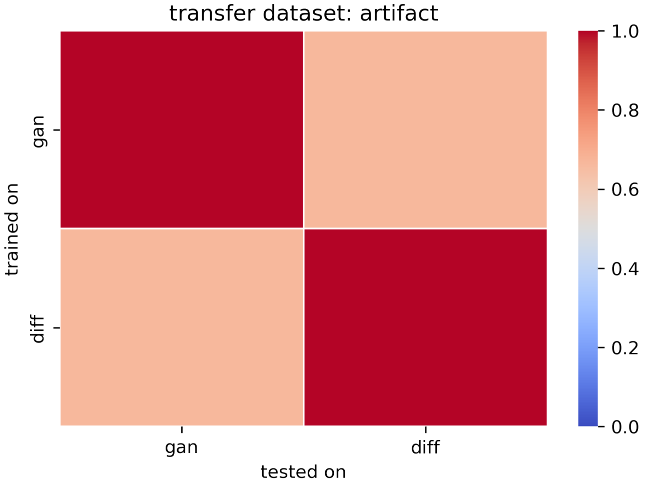


Figure 15: Limitation of the transferability. As described in section F, our experiment based on the ArtiFact consists of 8 clean datasets, 6 GAN, and 6 DM-generated images. The transferability is low, while the identification (see fig. 5) between clean and synthetic images is accurate.