Reconstruction-Based Membership Inference Attacks are Easier on Difficult Problems

Avital Shafran Shmuel Peleg Yedid Hoshen
The Hebrew University of Jerusalem
Jerusalem, Israel

Abstract

Membership inference attacks (MIA) try to detect if data samples were used to train a neural network model, e.g. to detect copyright abuses. We show that models with higher dimensional input and output are more vulnerable to MIA, and address in more detail models for image translation and semantic segmentation, including medical image segmentation. We show that reconstruction-errors can lead to very effective MIA attacks as they are indicative of memorization. Unfortunately, reconstruction error alone is less effective at discriminating between non-predictable images used in training and easy to predict images that were never seen before. To overcome this, we propose using a novel predictability error that can be computed for each sample, and its computation does not require a training set. Our membership error, obtained by subtracting the predictability error from the reconstruction error, is shown to achieve high MIA accuracy on an extensive number of benchmarks.

1. Introduction

Deep neural networks have been widely adopted in various computer vision tasks, e.g. image classification, semantic segmentation, image translation and generation etc. Due to the high sample-complexity of such models, they require large amounts of training data. However, obtaining many training samples might not be an easy task. In fact, collection and annotation is often an expensive and labor intensive process. In some domains, such as medical imaging, publicly available training data are particularly scarce due to privacy concerns. In such settings, a common solution is training the model privately and then providing black-box access to the trained model. However, even black-box access may leak sensitive information about the training data.

Membership inference attacks (MIA) are one way to detect such leakage. Given access to a data sample, an attacker attempts to find whether or not the sample was used in the training process.

Due to memorization in deep neural networks, prediction confidence tends to be higher for images used in training. This difference in prediction confidence helps MIA methods to successfully determine which images were used for training. Therefore, in addition to detecting information leakage, MIA also provide insights on the degree of memorization in the victim model.

MIA has previously been applied to a variety of neural network tasks including: classification, generative adversarial models, and segmentation. The accuracy achieved by MIA can vary greatly as a function of different properties of the attempted tasks. Our empirical results highlight two properties that make tasks more vulnerable to MIA attacks: i) Uncertainty: tasks where there is more uncertainty in the prediction of the output given an input are more susceptible to MIA. ii) Output dimensionality: tasks with higher-dimensional outputs are more vulnerable to MIA.

Motivated by the above findings, we focus our attention on two tasks that exhibit these properties: supervised image translation and semantic segmentation, including medical image segmentation. We begin by evaluating a simple-to-implement but effective MIA that uses pixel-wise reconstruction error between the model output and ground truth. This approach exploits memorization of the training data in the victim model, resulting in lower reconstruction error on images used for training. However, we observe that for some sample images the ground truth result can be easily predicted, and for others it is harder to predict. Reconstruction error alone is therefore less accurate at discriminating between hard to predict samples used in training and easy samples not seen before. To overcome this limitation, we propose a novel predictability error which is computed for each query input image and its ground truth output.

Our predictability error uses the accuracy of a linear predictor computed over the given query image, predicting pixel values from deep features of the input image. The linear predictor serves as a simple approximation of the task attempted by the victim model, providing an indication of the ease of predicting the output image from the input. The reconstruction error, together with the predictability error, helps to

Model	Dataset	Reconstruction Error	Membership Error
Pix2pix	Facades	93.62%	**97.59%**
Pix2pix	Maps2sat	84.22%	**85.65%**
Pix2pix	Cityscapes	77.74%	**83.23%**
Pix2pixHD	Facades	98.92%	**99.95%**
Pix2pixHD	Maps2sat	95.74%	**99.42%**
Pix2pixHD	Cityscapes	96.04%	**99.09%**
SPADE	Cityscapes	99.75%	**100%**
SPADE	ADE20K	85.31%	**89.79%**
UperNet50	ADE20K	96.80%	**98.09%**
UperNet101	ADE20K	95.74%	**96.94%**
HRNetV2	ADE20K	83.67%	**85.92%**
Inf-Net	COVID19	97.16%	**99.01%**
PraNet	Polyp	96.03%	**96.38%**

Table 1. Membership attack ROCAUC using our (i) reconstruction error L_{rec} and (ii) membership error L_{mem}. Using the membership error, which subtracts the image predictability error from the reconstruction error, substantially improves performance.

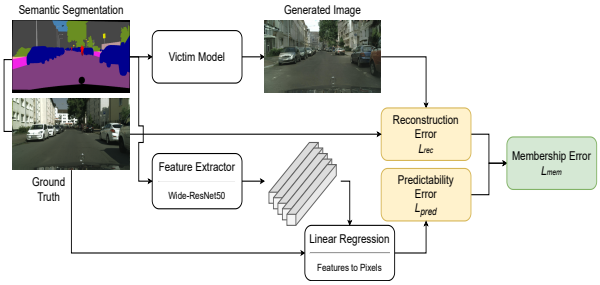


Figure 1. Illustration of the proposed black-box membership inference attack. Here shown for the case of image translation over the Cityscapes dataset. We would like to determine if a given sample was used in training. The victim model predicts a reconstructed image based on the input. In the top path the difference between the reconstructed image and the ground truth image gives the reconstruction error L_{rec}. In the bottom path we compute the predictability error L_{pred} of the sample from the error of a linear predictor to predict pixel values of the ground-truth image from deep features of the input. Subtracting L_{pred} from L_{rec} gives the membership error, L_{mem}.

discriminate between two factors of variation in the reconstruction error: (i) The "intrinsic" difficulty of the generation task for each image, based on its predictability error, and (ii) The boost in accuracy due to memorization of the training images. Defining a membership error that subtracts the predictability error from the reconstruction error is shown empirically to achieve high success rates in MIA.

Differently from other MIA approaches, we do not assume the existence of a large number of in-distribution data samples for training a shadow model - but rather operate on merely a single sample, using only a single query to the victim model. Our method is demonstrated to be effective over strong baselines on an extensive number of benchmarks, taken from image translation, semantic segmentation, and medical image segmentation. We discuss possible defenses against MIA and show their ineffectiveness against our method.

Our main contributions are:

1. Highlighting two key properties of tasks that are highly vulnerable to MIA.
2. Presenting the first MIA on image translation models.
3. Proposing a general single-sample, self-supervised, image predictability error for MIA.

2. Related Work

2.1. Membership Inference Attacks (MIA)

Shokri et al. [36] were the first to study MIA against classification models in a black-box setting. In this setting the attacker can only send queries to the victim model and get the full probability vector response, without being exposed to the model itself. They proposed to train multiple shadow models to mimic the behavior of the victim model, and then use those to train a binary classifier to distinguish between known samples from the train set and unknown samples. They assume the existence of in-distribution new training data and knowledge of the victim model architecture.

Salem et al. [35] further relaxed those assumptions and demonstrated that using only one shadow model is sufficient, and proposed using out-of-distribution dataset and different shadow model architectures, for a slightly inferior attack. Even more interestingly, they showed that without any training, a simple threshold on the victim model's confidence score is sufficient. This shows that classification models are more confident of samples that appeared in the training process, compared to unseen samples.

Sablayrolles et al. [34] proposed an attack based on applying a threshold over the loss value rather then the confidence and showed that black-box attacks are as good as white-box attacks. As the naive defense against such attacks is to modify the victim model's API to only output the predicted label, other works proposed label-only attacks [48, 27, 7].

While most previous work has been around classification models, there has been some effort regarding MIA on generative models such as GANs and VAEs [5, 19, 22]. An attack against semantic segmentation models was proposed by He et al. [21], where a shadow semantic segmentation model is trained, and is used to train a binary classifier. The classifier is trained on image patches, and final decision regarding the query image is set by aggregation of the per-patch classifi-

cation scores. The input to the classifier is a structured loss map between the shadow model's output and the ground truth segmentation map. Although this task is the closest to ours, our work is the first study of membership inference attacks on image translation models. We also note that [21] consider the setting where other input-output samples from the data distribution (or a very similar distribution) are available, whereas our attack does not require this information.

Besides membership inference attacks, other privacy attacks against neural networks exist. We refer the reader to sec. A.1 in the appendix for more details.

2.2. Conditional Image Generation

Image-to-image translation is the task of mapping an image from a source domain to a target domain, while preserving the semantic and geometric content of the input image. Currently, the most popular methods for training image-to-image translation models use Generative Adversarial Networks (GANs) [17] and are currently used in two main scenarios: (i) unsupervised image translation between domains [52, 26, 28, 6]; (ii) serving as a perceptual image loss function [23, 46, 53]. In this work we introduce the novel task of MIA on conditional image generation models.

2.3. Semantic Segmentation

Semantic segmentation is the task assigning a class label to each pixel in the input image. This can be thought of a classification problem for each pixel. State-of-the-art methods [47, 38] are based on fully convolutional networks and multi-scale representations of the input [29, 31].

3. Difficulty-based MIA

We investigate the effect of task difficulty and dimensionality on the success of MIA. Consequently, we concentrate on two promising tasks for MIA, image translation and semantic segmentation. We also present a novel image predictability error which significantly improves MIA accuracy.

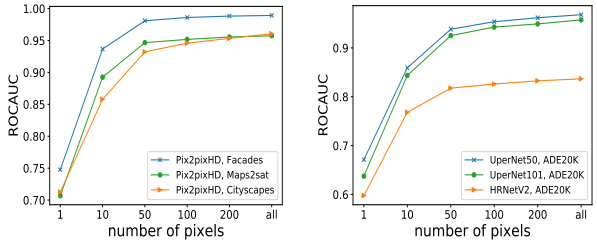
3.1. Effects of task difficulty and dimensionality

Every neural network model is a potential target for MIA. Previous work attempted MIA on many different models (classification, GANs, segmentation) with highly variable accuracy. In this section we present an investigation into two factors that affect MIA accuracy: task difficulty and output dimensionality. Full details are provided in Sec. 4.1. We perform reconstruction-based MIA by measuring the reconstruction error between the model output and the ground truth. This is done for multiple models, datasets and tasks. The attack method is described in more detail in Sec. 3.2.

MIA accuracy vs. task difficulty: We present results of reconstruction-based attack on three different tasks of different difficulties. We define task difficulty as the uncertainty in the output given the input image. The tasks are: i) auto-encoding - translating an image to itself. ii) Segmentation-to-image translation. iii) Landmark to face translation. The first task is the easiest as the output is trivially determined by the input. Landmark-to-face is harder than segmentation-to-image as the input contains less information on the output (e.g. no information on the identity of the face requires much more memorization). The results are presented in Tab. 2, where it can be seen that indeed MIA performance is more accurate for harder tasks.

Model	Task	Reconstruction Error
NVAE	CelebA2Self	50.74%
Pix2pixHD	Maps2sat	95.74%
Pix2pixHD	Cityscapes	96.04%
Pix2pixHD	Landmarks2CelebA	99.54%

Table 2. Comparison of reconstruction-based MIA accuracy on tasks with different difficulties. Easier tasks, e.g. CelebA2Self, in which there no uncertainty in the output given the input image, suffer less from memorization of the training data and therefore have lower vulnerability to MIA. As the uncertainty increases (segmentation maps and landmarks) models tend to memorize the training data and therefore the MIA accuracy increases.



(a) Pix2pixHD (b) Semantic Segmentation

Figure 2. Effect of reducing output dimensionality over a reconstruction-based attack. MIA accuracy is correlated with the output dimension, i.e. number of pixels, demonstrating that high output dimensionality tasks are more vulnerable to MIA.

MIA accuracy vs. output dimensionality: Many MIA approaches attack classification networks that have only a single output, usually a probability vector or in the more restrictive case, a single label. It is natural to ask if tasks with higher dimensional outputs are more vulnerable to MIA due to the ensemble effect of attacking each individual output dimension. In Fig. 2, we provide a comparison of reconstruction-based MIA when subsets of different sizes are used as the output. Note that segmentation with only a single pixel output is equivalent to classification. We can observe that MIA accuracy indeed scales with output dimensionality.

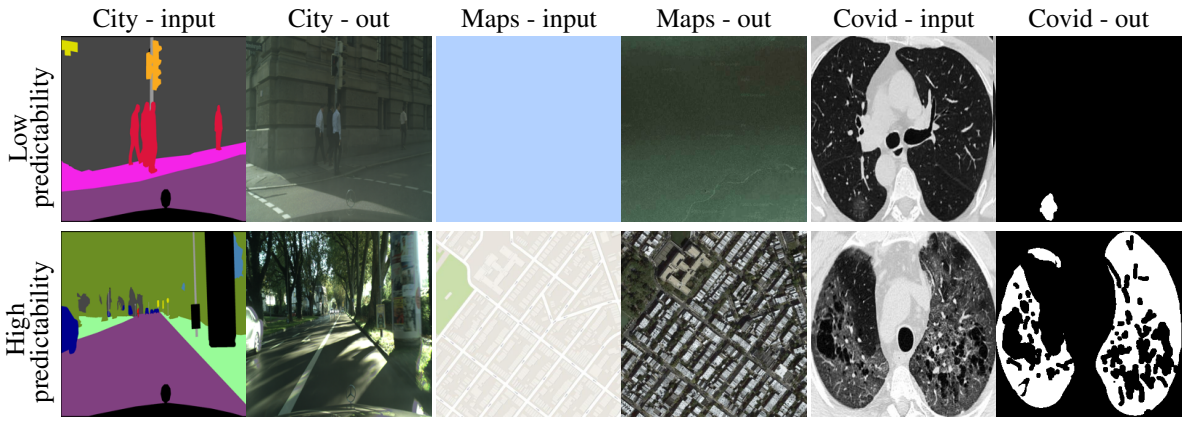


Figure 3. Examples of input-output pairs from the Cityscapes, Maps2sat and COVID19-CT datasets that received the lowest (first row) and highest (second row) predictability errors using our single-sample approach. It can be seen that detailed images with complicated patterns are ranked as difficult to predict, while images with less details and lower contrast are ranked as easier to predict.

3.2. MIA for high output dimensionality

We showed in Sec. 3.1 that both task difficulty and output dimensionality are correlated with the accuracy of MIA. We therefore focus on two important but difficult image tasks that have high-dimensional outputs: image translation and semantic segmentation. To the best of our knowledge, this is the first paper to consider MIA on image translation models.

We propose a simple-to-implement but effective attack. Our attack assumes a restrictive setting, where the attacker only has black-box access to the victim model $\mathbf{V}$. Differently from most previous works , we do not use shadow models or train a binary classifier, and thus do not require any additional training data and query the victim model only once.

Our membership attack is performed on a pair of query images (x, y) where $x \in \mathbb{R}^{h \times w}$ is an image from the input domain (h and w are the image height and width respectively) and $y \in \mathbb{R}^{h \times w}$ is the ground truth from the output domain. The requirement of the availability of the ground truth image y is in-line with previous works, and is a reasonable assumption in our target scenario. For each query we compute a membership error, L_{mem} (see Eq. (2)), to which we apply a pre-defined threshold τ, such that all queries where $L_{mem}(x, y) < \tau$ are marked as members of the training data. The membership error has two elements: reconstruction error and predictability error.

3.2.1 Reconstruction Error for Membership

Typical MIA on classification models consider the probability (or confidence) given by the model to the correct class. Semantic segmentation is an extension where the output is a probability vector for each pixel. Image translation models are different as they output a color value of each pixel. This value is the maximum likelihood estimate, and no probability distribution over possible values is given.

We propose to use the loss term as a reconstruction error, L_{rec}, to compute the pixel-wise difference between the output produced by a black-box access to the model, $\mathbf{V}(x)$, and the ground truth y. For semantic segmentation, where the output is a probability vector for each pixel, we use the cross-entropy error. For medical segmentation, we use the weighted IoU (Intersection over Union) loss and binary cross entropy loss. In the case of image-translation we use the L_1 error as we do not assume access to the discriminator and therefore can not use a GAN loss.

Due to memorization during the training process, the model output typically has lower reconstruction errors for images in the training set compared to unknown images.

3.2.2 Predictable and Unpredictable Images

In this section we address the following question: Given an input-output sample, is the output easily predictable from the input. Consider, for example, supervised segmentation-to-image translation. I.e., the task is to "invert" the segmentation process, and recover the original image that gave rise to a given segmentation map. It is clear that not all cases are equally predictable: (i) hard to predict images have sharp and detailed textures, whereas more predictable images have blurrier textures; (ii) images with semantic segmentation maps that contain only few categories provide less guidance than those with more detailed segmentation maps, making the correct prediction less certain. The image predictability error should quantify these observations. In Sec. 4.2 we show that such a predictability error is important for increasing accuracy of membership inference attacks.

We briefly describe two previous approaches for measuring image difficulty:

Human-Supervised: Ionescu et al. [41] proposed to de-

fine image difficulty as the human response time for solving a visual search task. For this, they collected human annotations for the PASCAL VOC 2012 dataset [10] and trained a regression model, based on pre-trained deep features, to predict the collected difficulty score. The disadvantage of this method is that human-specified difficulty scores may not correlate to the predictability of the image synthesis by neural networks. This is demonstrated empirically in Sec. 4.2.1.

Multi-Image: Another approach taken by Chen et al. [5] is training a model on a set of image pairs similar to the target distribution. This approach uses the reconstruction error of the external model on the target image pairs as its predictability error - larger reconstruction errors correspond to harder to predict images. This approach has a significant drawback: a large number of images, similar to the target image, are required in order to learn a reliable generative model. In many cases, images from the target distribution may not be available. Additionally, training a model for every task is tedious and computationally expensive.

Proposed - Single-sample predictability error: We propose a novel method to assign a predictability error for models with image outputs. The predictability error measures the success of a linear regression model to predict output pixel values from a high-level representation of the input image.

A related approach was proposed by Hacohen et al. [18] for measuring image difficulty for classification models. Our method is significantly different as it is trained on a single input-output sample rather than on a large dataset.

The linear regression model uses image features of a pre-trained Wide-ResNet50×2 [50]. We concatenate the activation values in the first 4 blocks, giving 56×56 feature vectors of size 3840 each. The output image resolution is reduced to 56×56 to match the size of the first Wide-ResNet50×2 block. We denote the concatenated feature vector for pixel i as $\psi(i)$.

The linear regression model $\mathbf{P}$ is a matrix of size 3840×3, multiplied with the feature vector $\psi(i)$ of pixel i to give an estimate of the RGB colors y^i. We optimize $\mathbf{P}$ over 70% randomly selected pixels. The image predictability error is the average absolute error over the 30% unselected pixels:

$$L_{pred}(x,y) = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{P}\psi(i) - y^i\|_1 \qquad (1)$$

where y^i is the ground truth value of the i_{th} pixel in the resized ground truth image y. Fig. 3 presents some images that received the highest and lowest predictability errors. See sec. A.2 for more details.

3.2.3 Membership Error

As observed before, for some samples the output images can be easily predicted from the input, while for other samples the output can not be predicted. While the reconstruction error achieves high MIA success rates, it has a significant limitation - it does not discriminate between predictable and unpredictable samples. The victim model will have higher errors when generating unpredictable training samples, and lower errors when generating easily predictable ones. In such samples reconstruction error may result in wrong membership classification.

Given our image predictability error L_{pred} in Eq. (1) and the reconstruction error L_{rec} we calculate a membership error L_{mem} as follows:

$$L_{mem}(x,y) = L_{rec}(x,y) - \alpha \cdot L_{pred}(x,y) \qquad (2)$$

L_{mem} is computed by subtracting the predictability error L_{pred} from the reconstruction error L_{rec} weighted by α. Unless specified otherwise, we use $\alpha = 1.0$, and present the effect of different α values in sec A.3. This lowers the membership error L_{mem} for harder-to-predict images compared to easier-to-predict images having the same reconstruction error. See Fig. 1 for an overview illustration of our method.

Using the membership error L_{mem} (2) for MIA substantially improves the success rates in all of our experiments, as shown in Tab. 1 and Fig. 4

4. Experiments

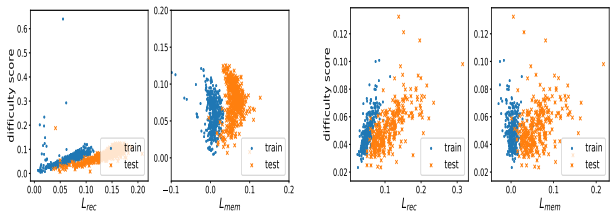
We first investigate two factors that affect MIA accuracy, task difficulty and output dimensionality, and show that MIA attacks are easier on difficult tasks with high output dimensionality. We then extensively evaluate our MIA on image translation and semantic segmentation networks. We also compare our single-sample predictability error against strong baselines. Additional results and ablations can be found in the appendix. In accordance with previous membership attack works, the success rate is measured using the area under the ROC curve (ROCAUC) metric.

4.1. Effectiveness of membership inference attacks

As mentioned in Sec. 3.1, there has been extensive research on MIA against various neural networks, resulting in variable accuracy. We investigated two factors that affect MIA accuracy: task difficulty and output dimensionality.

4.1.1 MIA accuracy vs. task difficulty

We defined task difficulty as the uncertainty of the output given the input image. In the limit of sufficiently large training datasets, when models are trained to perform easy tasks, such as auto-encoding - translating an image to itself, they are able to generalize well to unseen images, and do not need to depend on memorization of the training data in order to minimize the loss function. As membership inference attacks are highly correlated with model memorization, their



(a) Pix2pixHD - Maps (b) Pix2pixHD - Cityscapes

Figure 4. The proposed membership error L_{mem} can better separate train (blue) and test (orange) images by a simple threshold (i.e. a vertical line) compared to the reconstruction error L_{rec}.

performance decreases on such tasks. Similarly, models struggle with learning difficult tasks, in which the input does not contain sufficient information to full specify the output, and therefore the loss minimization encourages the model to memorize the training samples. This lack of predictability acts as a strong motivation for memorization, even at the limit of well-trained models trained on large datasets, provided sufficient capacity.

In order to demonstrate this, we performed reconstruction-based MIA, by using L_{rec} described in Sec. 3.2.1, on three tasks of different levels of difficulty. The first and easiest task is auto-encoding. For this, we attack a NVAE [43] model, trained on the CelebA [30] dataset. The second task, more difficult than auto-encoding, is segmentation-to-image translation. We attack two Pix2PixHD models [46], trained on the Maps2sat [23] and Cityscapes [8] datasets. The third and most difficult task is landmarks-to-face translation. For this task we extracted facial landmarks [3] from 50K CelebA images [30]. We consider this to be the most difficult task out of the three as the input contains less information on the output in comparison to segmentation maps (e.g. no information regarding the identity of the face). Results, presented in Tab. 2, demonstrate that reconstruction-based MIA are more successful on difficult tasks.

4.1.2 MIA accuracy vs. output dimensionality

Previous works mostly focused on MIA against classification models, where there is a single output, i.e. probability vector or in the more restrictive case, a single label. It is natural to ask whether higher dimensional outputs are more vulnerable due to the ensemble effect of combining the attacks on each individual output dimension to a stronger, joint attack.

We perform reconstruction-based MIA on the Pix2PixHD architecture, trained on the CMP Facades [42], Maps2Sat and Cityscapes datasets, as well as on three semantic segmentaion models - UperNet50, UperNet101 [47] (using ResNet50 and ResNet101 as backbones) and HRNetV2 [38] - trained on the ADE20K dataset [51]. Fig. 2 demonstrates the effect of reducing the output dimension on the attack accuracy. The reduction is achieved by randomly sampling N output pixels, and using them as the output, where N ranges from a single pixel and up to 200 pixels. Note that in the case of semantic segmentation, having only a single pixel output is equivalent to classification. We repeat this experiment 10 times and report the average result.

We observed that MIA accuracy indeed scales with the number of output dimensions. Results for other models are presented in Fig. 7.

4.2. Membership inference attack accuracy evaluation

We evaluate our membership attack on three image translation architectures, Pix2Pix [23], Pix2PixHD [46], SPADE [33], three semantic segmentation architectures - UperNet50 and UperNet101 [47] (ResNet50 and ResNet101 as backbones), HRNetV2 [38] as well as two medical segmentation architectures - Inf-Net [12] and PraNet [11]. We evaluated on various datasets, including CMP Facades [42], Maps2sat [23], Cityscapes [8], ADE20K [51].

In the case of medical segmentation we evaluated two tasks: lung infection segmentation from Covid-19 CT images [12] and polyp segmentation in colonoscopy images [37, 2, 39, 44, 25].

All pix2pix and pix2pixHD models are trained from scratch, with the exception of the Cityscapes dataset on the Pix2PixHD architecture in which we use the supplied large pre-trained model for computational constraints on the high resolution. The rest of the models are pre-trained.

It can be seen in Tab. 1 that while using the reconstruction error alone achieves a high success rate, the membership error (which calibrates the result by sample predictability) significantly improves the results. Fig 4 demonstrates the effect of subtracting the predictability error from the reconstruction error. After calibration, a single threshold on the membership error can separate train and test images. Results on additional benchmarks are presented in Fig. 8 in the appendix.

Utilizing common image augmentations, i.e. horizontal flipping and random cropping, in order to construct a larger a larger set $\{(x_{aug}, y_{aug})\}$ has a small impact over the attack accuracy, as the output dimension is large enough as is.

4.2.1 Comparison to Human Supervision

We compare our self-supervised single-sample predictability error with the human-supervised difficulty score described in Sec. 3.2.2. This score was proposed by Ionescu et al. [41], which defined image difficulty to be the human response time for solving a visual search task. In order to provide a fair comparison, we replace the pretrained VGG-f [4] features, used by [41], with the more powerful pretrained

Model	Dataset	Single		Multi
		Ours	Superv.	In-Dist.
Pix2pix	Facades	**97.59%**	93.67%	-
Pix2pix	Maps2sat	85.65%	86.48%	**92.43%**
Pix2pix	Cityscapes	**83.23%**	77.06%	82.47%
Pix2pixHD	Facades	**99.95%**	98.86%	-
Pix2pixHD	Maps2sat	**99.42%**	98.38%	82.87%
Pix2pixHD	Cityscapes	**99.09%**	96.86%	94.76%
UperNet50	ADE20K	**98.09%**	96.79%	79.47%
UperNet101	ADE20K	**96.94%**	95.49%	76.01%
HRNetV2	ADE20K	**85.92%**	84.42%	84.48%

Table 3. MIA accuracy of our self-supervised single-sample method vs. using human-supervised single-sample and multi-image baselines for the predictability error. Note that in-distribution multi-image requires extra supervision of 100 images

Model	Dataset	Ours	Shadow Model
Pix2pix	Maps2sat	**85.65%**	80.15%
Pix2pix	Cityscapes	**83.23%**	78.68%
Pix2pixHD	Maps2sat	**99.42%**	98.63%
Pix2pixHD	Cityscapes	**99.09%**	96.39%

Table 4. Comparison between our MIA and the popular shadow-model-based classifier attack, using 100 train and 100 test samples. Our MIA outperforms while not requiring extra training images.

Wide-ResNet50 $\times$ 2 [50] features we use in our predictability error. Samples of images ranked as easy and hard by the supervised score are presented in Fig. 9 in the appendix. As can be seen in Tab. 3, our self-supervised single-sample predictability error outperforms the human-supervised difficulty score. In sec. A.6, we provide a comparison of the relation between the reconstruction error and the supervised score to the relation between the reconstruction error and our self-supervised predictability error, and show that our score is better correlated to the reconstruction error.

4.2.2 Comparison to Multi-Image Scores

Although our MIA method does not require the availability of multiple auxiliary samples from the target distribution or from a similar distribution, it is interesting to compare our single sample predictability error to methods that use multiple samples. We compute multi-sample predictability errors (MSPE) by training a "shadow" model to map the input to output images in the auxiliary samples. As an upper-bound on MSPE performance, the shadow model is given exactly the same architecture as used by the victim model (although this knowledge may not be available in practice). The MSPE is defined by the reconstruction error of the shadow model on the target sample. Two scenario were evaluated:

In-distribution data: In this setting the shadow model's data shares the distribution of the victim's training data, by being trained on 100 randomly sampled image pairs from the test set of the corresponding dataset. Facades was not used as it did not have enough images. The results are presented in Tab. 3. For Pix2PixHD and semantic segmentation, MSPE underperformed our method (as 100 samples are insufficient for training such large models). As Pix2Pix is a smaller network, MSPE was more successful there, obtaining competitive results with our method. Note that it still requires extra samples, often not available. We analyzed the number of samples required for MSPE to reach the accuracy of our method, in most tasks, even 100 were insufficient (see Fig. 10).

Auxiliary dataset: As suggested by He et al. [21], we also compare our method to the setting were many out-of-distribution but similar samples are available. We trained shadow models on $4K$ image pairs from the BDD dataset [49] as MSPE for the Cityscapes dataset, as both datasets consist of street scene images and have compatible label spaces. We found that MSPE underpeformed our method by $10\% - 30\%$. (see sec. A.7 for exact results). Note that it is rare to have similar datasets with nearly identical labels. Cityscapes was the only dataset from those evaluated in this paper for which such a similar dataset could be found.

Shadow-model classifiers: Although deviating somewhat from predictability errors, for the interest of completeness, we report the ROCAUC results of the popular approach of shadow-model classifiers for image translation MIA, see Tab. 4 (classification accuracy is lower, see Tab. 9). We use the approach of He et al. [21] and train a classifier to distinguish between the "loss maps" of the train and test auxiliary samples of the shadow model. The classifier is then used to score images of the target dataset as train or test (see [21] for the complete details).

We show that this approach underperforms our method in both in-distribution and auxiliary dataset settings (exact results presented in Tab. 9). It is surprising that shadow models do not perform well on image translation MIA as they are very effective for image segmentation MIA (as shown in [21]). We believe the difference in performance can be explained by the fact that segmentation maps have similar distributions between datasets with similar label spaces while natural images have very different distribution - making membership classifiers on the image2seg task generalize better than the seg2image task. We note again, that such techniques requires the availability of auxiliary in-distribution samples or very similar datasets which is often not possible. For example, medical segmentation datasets are often quite small due to the sensitivity of the data and high cost of obtaining ground truth labels. The Covid-19 CT dataset [12] is composed of 50 train and 50 test samples, too small to apply a shadow model based attack.

Model	Dataset	Orig	No L_{rec}
Pix2pix	Maps2sat	84.22%	68.44%
Pix2pix	Cityscapes	77.74%	51.06%
Pix2pixHD	Maps2sat	95.74%	75.76%
Pix2pixHD	Cityscapes	96.04%	56.58%

Table 5. Effect of cGAN and reconstruction losses on the accuracy of reconstruction-based MIA. The cGAN loss is less susceptible to MIA.

Model	Dataset	No-defense	Argmax-defense
UperNet50	ADE20K	98.09%	98.05%
UperNet101	ADE20K	96.94%	97.11%
HRNetV2	ADE20K	85.92%	89.32%

Table 6. Comparison between our attack accuracy (ROCAUC) on undefended semantic segmentation models and models defended by the argmax defense. As can be seen, our attack still manages to succeed much better than random guessing.

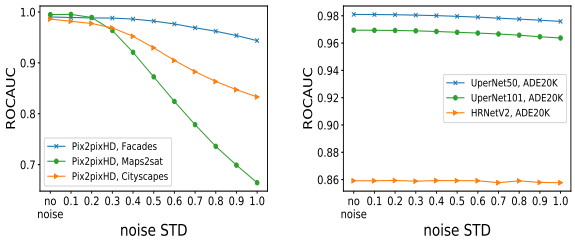
5. Discussion

Effect of Memorization: Membership inference attacks are closely related to memorization in the victim model. In order to better understand this relation, we measure the success of our attack under different levels of memorization. We do so by evaluating our attack on checkpoints saved at different epochs during the training of our image-translation victim models. We observed that as the training process progresses, the victim model memorizes the training data which results in higher attack success rates (See sec. A.9).

Reconstruction loss effect on MIA: We evaluated the effect of the reconstruction loss term on the accuracy of reconstruction-based MIA. For this, we compared the accuracy of the attack against image translation models, trained using both reconstruction and cGAN loss terms versus models that were only trained using the cGAN loss term. As can be seen in Tab. 5, the reconstruction loss indeed has a significant impact over the attack accuracy.

Argmax defense: In the argmax defense, the victim model returns only the predicted label, rather then the full probability vector. As image translation models predict pixel values, and does not output probability vectors, this defense does not apply. For semantic segmentation models, we evaluate our attack against this defense, by replacing the cross-entropy loss in L_{rec} to the L_0 error. As can be seen in Tab. 6, the attack efficacy is not reduced, demonstrating the weakness of this defense.

Differential private SGD (DP-SGD) defense: In the defense by Abadi et al. [1], the commonly used Stochastic Gradient Descent optimization algorithm is modified in order to provide a differentially private [9] model. This is done by adding Gaussian noise to clipped gradients for each sample in every training batch. There exists a trade-off between privacy and utility, in which the amount of added noise must be large enough to ensure privacy while not degrading the model's outputs to the point where the model is useless. Training a deep model with DP-SGD is an unstable process. We experimented with multiple common configurations, i.e. added noise ratios and maximal gradient clipping threshold, and were not able to find a configuration that yields visually satisfying results. Hence, although the DP-SGD defense is theoretically protecting the victim model against membership inference attacks, in practice we find it to be impractical against our attack as it results with total corruption of the victim model.



(a) Pix2pixHD (b) Semantic Segmentaion

Figure 5. Effect of Gauss defense on the attack accuracy. Even with large amounts of added noise, our attack still manages to succeed much better then random guessing.

Gauss defense: In this defense, we add Gaussian noise to the image generated by the victim model [16]. This attempts to hide specific artifacts of the model. We evaluate our attack accuracy as a function of different noise STD. Fig. 5 shows that a considerable amount of noise, which corrupts the generated output, is required in order to have a significant effect over our attack success. Moreover, it can be seen that even with large amounts of noise, our attack still manages to succeed much better than random guessing. This implies that our attack is robust to the Gauss defense. Results on additional benchmarks are presented in Fig. 12.

6. Conclusion

In this work, we highlight two properties that make tasks more vulnerable to MIA: i) Uncertainty: tasks where there is more uncertainty in the prediction of the output given an input ii) Output dimensionality: tasks with high-dimensional output. We show that a black-box reconstruction-based membership attack is very effective on two tasks that exhibit these properties: image translation and semantic segmentation, including medical segmentation. We further improve the attack by proposing a novel image predictability error. Our membership error, composed of the reconstruction and predictability errors, has been extensively evaluated on various benchmarks and was shown to achieve high accuracy.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

[4] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv:1405.3531*, 2014.

[5] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against gans. *arXiv preprint arXiv:1909.03935*, 2019.

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018.

[7] Christopher A Choquette Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. *arXiv preprint arXiv:2007.14321*, 2020.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[9] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. of Computer Vision*, 88(2):303–338, June 2010.

[11] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. *MICCAI*, 2020.

[12] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE TMI*, 2020.

[13] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

[14] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 17–32, 2014.

[15] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 619–633, 2018.

[16] Justin Gilmer, Nicolas Ford, Nicholas Carlini, and Ekin Cubuk. Adversarial examples are a natural consequence of test error in noise. In *ICML*, pages 2280–2289, 2019.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[18] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *ICML*, pages 2535–2544, 2019.

[19] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(1):133–152, 2019.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[21] Yang He, Shadi Rahimian, Bernt Schiele, and Mario Fritz. Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation. *arXiv preprint arXiv:1912.09685*, 2019.

[22] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(4):232–249, 2019.

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[24] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020.

[25] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[26] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, pages 1857–1865, 2017.

[27] Zheng Li and Yang Zhang. Label-leaks: Membership inference attack with label. *arXiv preprint arXiv:2007.15528*, 2020.

[28] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[29] Xiaolong Liu, Zhidong Deng, and Yuhan Yang. Recent progress in semantic image segmentation. *Artificial Intelligence Review*, 52(2):1089–1106, 2019.

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, December 2015.

[31] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *arXiv:2001.05566*, 2020.

[32] Seong Joon Oh, Bernt Schiele, and Mario Fritz. Towards reverse-engineering black-box neural networks. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 121–144. Springer, 2019.

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019.

[34] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *ICML*, pages 5558–5567, 2019.

[35] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

[36] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[37] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014.

[38] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv:1904.04514*, 2019.

[39] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

[40] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 601–618, 2016.

[41] Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Popescu, Dim P Papadopoulos, and Vittorio Ferrari. How hard can it be? estimating the difficulty of visual search in an image. In *CVPR*, pages 2157–2166, 2016.

[42] Radim Tyleček and Radim Šára. Spatial pattern templates for recognition of objects with regular structure. In *Proc. German Conference on Pattern Recognition (GCPR)*, Saarbrucken, Germany, 2013.

[43] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *arXiv:2007.03898*, 2020.

[44] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017, 2017.

[45] Binghui Wang and Neil Zhenqiang Gong. Stealing hyperparameters in machine learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 36–52. IEEE, 2018.

[46] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pages 8798–8807, 2018.

[47] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018.

[48] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

[49] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv:1805.04687*, 2018.

[50] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv:1605.07146*, 2016.

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.

[53] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017.

A. Appendix

A.1. Other Privacy Attacks

Besides membership inference attacks, there exists a wide range of privacy attacks against neural networks. Model inversion attacks, first proposed by [14], aim at reconstructing features of the training data, e.g. recovering an image of a person from face recognition models [13]. Property inference attacks, proposed by [15], do not focus on the privacy of individual data samples, as in membership inference and model inversion attacks, but focus at inferring global properties of the training data, such as the environment in which the data was produced or the fraction of the data that comes from a certain class.

Model extraction attacks, also referred to as model stealing, attack a model f by constructing a substitute model $\hat{f}$ that is either identical or equivalent to f [40, 24]. Related line of work [45, 32] attempts to infer hyperparameters such as the optimization proccess, e.g. SGD or ADAM.

A.2. Detailed Description of Our MIA Algorithm

Our MIA consist of computing the two terms in Eq. (2), i.e. L_{rec} and L_{pred} for a given query pair (x, y), where x is an image from the input domain and y is the ground truth from the output domain, using only a black-box access to the victim conditional generation model $\mathbf{V}$.

L_{rec} is computed using the pixel-wise error between the output image predicted by the model, $\mathbf{V}(x)$, and the ground truth image y, see step 1 in the Algorithm 1. For image translation models, we set the pixel-wise error function, err, to be the L_1 loss:

$$L_{rec}^{trans}(x, y) = \|\mathbf{V}(x) - y\| \tag{3}$$

For semantic segmentation, where the output values are probability vectors rather then pixel values, we use the cross-entropy loss:

$$L_{rec}^{seg}(x, y) = -log(\mathbf{V}(x)[y]) \tag{4}$$

In the case of medical segmentation, following Fan et al. [11, 12], we use the weighted IoU loss and binary cross-entropy loss:

$$L_{rec}^{med}(x, y) = L_{IoU}^{w}(x, y) + L_{BCE}^{w}(x, y) \tag{5}$$

Defined as:

$$L_{IoU}^{w} = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} w_{ij}(\mathbf{V}(x)_{ij} \cdot y_{ij})}{\sum_{i=1}^{H}\sum_{j=1}^{W} w_{ij}(\mathbf{V}(x)_{ij} + y_{ij} - \mathbf{V}(x)_{ij} \cdot y_{ij})} \tag{6}$$

$$L_{BCE}^{w} = -\frac{\sum_{i=1}^{H}\sum_{j=1}^{W} w_{ij} log(\mathbf{V}(x)[y]_{ij})}{\sum_{i=1}^{H}\sum_{j=1}^{W} w_{ij}} \tag{7}$$

Where H and W are the height and width of the query sample, and w_{ij} is the weight of pixel (i, j) and is defined as follows, where A_{ij} represents the area that surrounds the pixel (i, j):

$$w_{ij} = 1 - \left| \frac{\sum_{m,n \in A_{ij}} y_{mn}}{\sum_{m,n \in A_{ij}} 1} - y_{ij} \right| \tag{8}$$

L_{pred} is computed as the average error of a linear regression model, $\mathbf{P}$, in predicting pixel values from deep features of the input image.

Our deep features are the activation values in the first 4 blocks of a pre-trained Wide-ResNet50×2 [50]. These features are of sizes $56\times56\times256$, $28\times28\times512$, $14\times14\times1024$, and $7\times7\times2048$. We interpolate all features to size 56×56 using bi-linear interpolation (step 2), and also reduce the output image to 56×56 using bicubic interpolation (step 3). This gives a concatenated feature vector of size 3840 for each pixel i in the 56×56 image ($256+512+1024+2048=3840$). We denote the concatenated feature vector for pixel i as $\psi(i)$.

We randomly select 70% of the pixels as train set, and compute a linear model $\mathbf{P}$ to estimate the RGB pixel values y_{train}^{i} from the corresponding deep features $\psi_{train}(i)$ (step 4). The remaining 30% of pixels will be used as a test split, $\{\psi_{test}, y_{test}\}$ (step 5). I.e. $|\psi_{train}| = 2195\times3840$, $|y_{train}| = 2195\times3$ and $|\psi_{test}| = 941\times3840$, $|y_{test}| = 941\times3$.

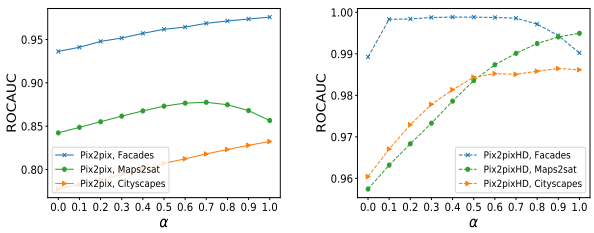
The linear regression model $\mathbf{P}$, a matrix of size 3840×3, is trained to minimize the error over $\{\psi_{train}, y_{train}\}$ (step 6). L_{pred} will be the average absolute error over $\{\psi_{test}, y_{test}\}$ (step 7). We found that fitting the linear model to 70% of pixels and measuring the error on the remaining 30% gives better results than just measuring the error of the linear fitting.

We compute L_{mem} according to Eq. (2) and compare the results with a predefined threshold value τ, such that any pair (x, y) for which is holds that $L_{mem}(x, y) < \tau$ is denoted as a member of the victim models' $\mathbf{V}$ train set (steps 8-9).

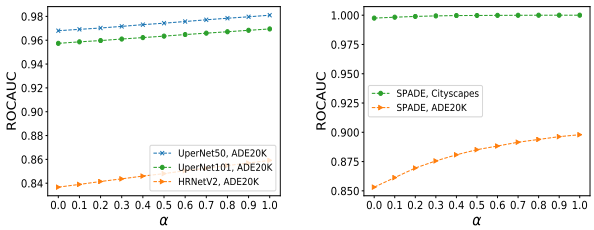
We have experimented with different resize methods (step 3) and found that our attack success rate is not very sensitive to the resize method. Additionally, we evaluated the effect of different train-test partitions (steps 4 & 5) and found that using less than 50% of the image pixels for training the linear regression model results with unstable performance, while all values of 50% or above results in similar attack success rates.

A.3. Parameter Selection

We experimented with different values for the α value in Eq. (2). As can be seen in Fig. 6, $\alpha = 1$ was the best choice over all benchmarks.



(a) Pix2pix (b) Pix2pixHD



(c) Semantic Segmentation (d) SPADE

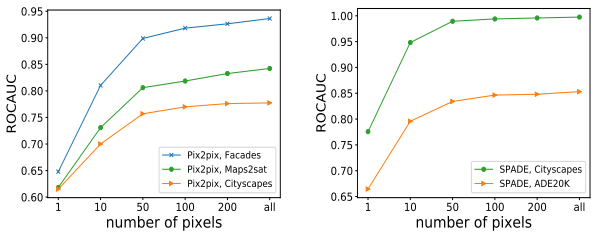
Figure 6. Effect of α in Eq. (2) over the attack success.

Algorithm 1. Membership Inference Attack
Input: Query pair (x, y), victim model $\mathbf{V}$, feature extractor $\mathbf{F}$, scalar α, threshold τ, error function err
Output: Membership inference result

1. $L_{rec} = err(\mathbf{V}(x), y)$
2. $\psi = \mathbf{F}(x) // \ |\psi| = 56 \times 56 \times 3840$
3. $y = resize(y, 56 \times 56 \times 3)$
4. $\{\psi_{train}, y_{train}\} \xleftarrow{70\%} \{\psi, y\}$
5. $\{\psi_{test}, y_{test}\} = \{\psi, y\} \setminus \{\psi_{train}, y_{train}\}$
6. Train linear regression $\mathbf{P}$ with $\{\psi_{train}, y_{train}\}$
7. $L_{pred} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{P}\psi_{test}(i) - y^i_{test}\|_1 \ // N = 941$
8. $L_{mem} = L_{rec} - \alpha \cdot L_{pred}$
9. **if** $L_{mem} < \tau$ **then**
   Return **True**
   else
   Return **False**

A.4. MIA vs output dimension

As described in Sec. 4.1, we evaluted the effect of reducing the output dimension on the accuracy of reconstruction-based MIA. The reduction was achieved by randomly sampling N output pixels, and using them as the output, where N ranges from a single pixel and up to 200 pixels. Fig. 7 demonstrates that MIA accuracy indeed scales with the number of output dimensions. Results for Pix2PixHD, UperNet and HRNetV2 are presetned in Fig. 2.



(a) Pix2pix (b) SPADE

Figure 7. Effect of reducing output dimensionality over a reconstruction-based attack. MIA accuracy is correlated with the decrease of output dimension, i.e. number of pixels, demonstrating that high output dimensionality problems are more vulnerable to MIA.

A.5. calibration Effect

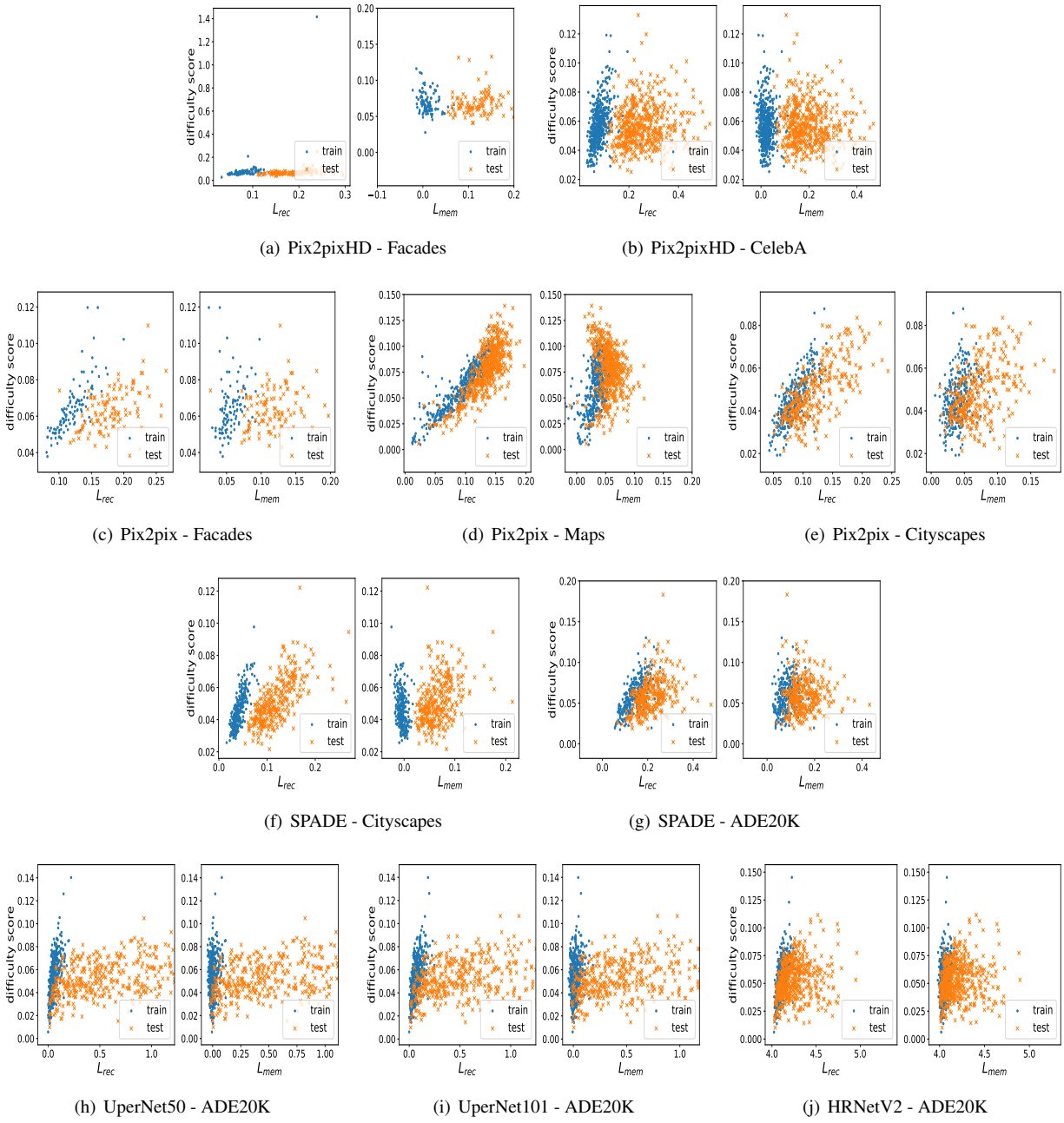
As can be seen in Tab. 1, using our membership error L_{mem}, Eq. (2), substantially improves the sucess rates in all of our experiments. As can be seen in Fig. 8, our L_{mem} can better separate train and test images by a simple threshold compared to the reconstruction error L_{rec}. Results for Pix2PixHD on the Maps2sat and Cityscapes datasets are presented in Fig. 4.

A.6. Human-Supervised Image difficulty score

We compare our self-supervised single-sample predictability error with the human-supervised difficulty score proposed by [41]. In Fig. 9, we present images ranked from easy to difficult using our implementation of the supervised-image difficulty score, for the Cityscapes and Maps datasets. The ranking seems correlated with image sharpness and level-of-detail images. As can be seen in Tab. 3, our score outperforms the human-supervised score. We compare the correlation between the reconstruction error for unseen images to our self-supervised predictability error and the human-supervsied scorre.

A.7. Multi-Image predictability error

As discussed in Sec. 4.2.2, we compare our single-sample predictability error to a multi-sample predictability error (MSPE) by training a "shadow" model, sharing the same architecture as the victim model, on auxiliary samples. As can be seen in Tab. 3, when training the MSPE on 100 images, it underperforms our method on Pix2PixHD and the evaluated semantic segmentation models. For the smaller Pix2Pix architecture, MSPE was more successful, obtaining



(a) Pix2pixHD - Facades (b) Pix2pixHD - CelebA

(c) Pix2pix - Facades (d) Pix2pix - Maps (e) Pix2pix - Cityscapes

(f) SPADE - Cityscapes (g) SPADE - ADE20K

(h) UperNet50 - ADE20K (i) UperNet101 - ADE20K (j) HRNetV2 - ADE20K

Figure 8. The proposed membership error L_{mem} can better separate train and test images by a simple threshold (i.e. a vertical line) compared to the reconstruction error L_{rec}. Pix2pixHD for Maps2sat and Cityscapes are presented in Fig. 4

competitive results with our method. We analyzed the effect of number of samples over the MSPE performance. As can be seen in Fig. 10, in most tasks, increasing the number of samples did not improve performance.

We also compare our method to the setting were many out-of-distribution but similar sample are available. We trained shadow models on $4K$ samples from the BDD dataset as MSPE for the Cityscapes dataset. As can be seen in Tab. 8, this too underperforms our method. Note that it is rare to have similar datasets with nearly identical labels, such as in the case of BDD and Cityscapes.

A.8. Shadow models

As discussed in Sec. 4.2.2, for the interest of completeness we compare our method with the popular approach of shadow-model classifiers for image translation MIA. For this,

Model	Dataset	Ours train / test	Human-Supervised train / test
Pix2Pix	Facades	0.79 / 0.50	-0.02 / 0.16
Pix2Pix	Maps2sat	0.51 / 0.77	0.79 / 0.52
Pix2Pix	Cityscapes	0.78 / 0.71	0.04 / 0.09
Pix2PixHD	Facades	0.67 / 0.36	0.27 / 0.04
Pix2PixHD	Maps2sat	0.38 / 0.79	0.77 / 0.56
Pix2PixHD	Cityscapes	0.76 / 0.62	0.36 / 0.48
SPADE	Cityscapes	0.80 / 0.68	0.29 / 0.53
SPADE	ADE20K	0.48 / 0.27	0.25 / -0.05
UperNet50	ADE20K	0.66 / 0.13	0.34 / 0.05
UperNet101	ADE20K	0.65 / 0.13	0.38 / 0.06
HRNetV2	ADE20K	0.61 / 0.22	0.36 / 0.10

Table 7. Our self-supervised difficulty score is better correlated with the reconstruction error than the human-supervised

Figure 9. Examples of images from the Cityscapes (first two rows) and Maps2sat (last two rows) datasets that received the lowest (first and third row) and highest (second and last row) predictability errors using the supervised difficulty score.

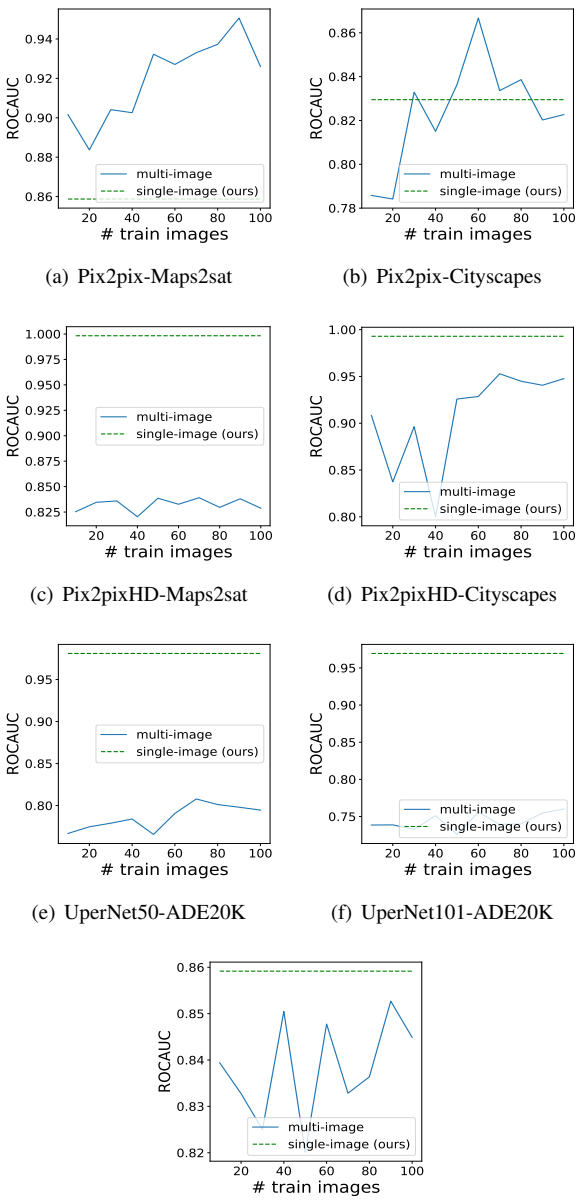
Model	Dataset	Ours	Multi-Image	
			In-Dist	BDD
Pix2pix	Cityscapes	**82.94%**	**82.47**%	74.43%
Pix2pixHD	Cityscapes	**99.29%**	96.86%	66.2%

Table 8. Comparison between our single-image predictability error and two multi-image baselines, using in-distribution images and a larger amound of out-of-distribution images (BDD).

we select N images, denoted as *shadow_train*, for training the shadow model. As an upper-bound, the shadow model is given the exactly same architecture as used by the victim model. Another N images, not seen by the shadow model, are set to be *shadow_test*. We define a new label for each sample as follows:

$$label(x) = \begin{cases} 0, & \text{if } x \leftarrow shadow\_train \\ 1, & \text{if } x \leftarrow shadow\_test \end{cases} \quad (9)$$

The classifier $\mathbf{C}$ architecture and training procedure are similar to [21]. For each image, we compute the structured loss map between the ground-truth image and the generated image, using L_1 as the loss function. At every epoch we randomly crop 15 patches of size 90×90 from the structured loss map. We train a ResNet-50 [20] from scratch on the 90×90 patches, modified for binary classification. We use a batch size of 8, SGD optimizer, weight decay of $1e-2$, initial learning rate of 0.1 which reduces by a factor of 0.1 every 15 epochs. As previously mentioned, we do not evaluate this on the Facades dataset, due to its size.



(a) Pix2pix-Maps2sat (b) Pix2pix-Cityscapes

(c) Pix2pixHD-Maps2sat (d) Pix2pixHD-Cityscapes

(e) UperNet50-ADE20K (f) UperNet101-ADE20K

(g) HRNetV2-ADE20K

Figure 10. Comparison of MIA accuracy when using our single sample vs. using multi-sample predictability errors, as a function of the number of training images. Note that the multi-image score assumes knowledge of the victim's model, as well as the availability of many labeled training images

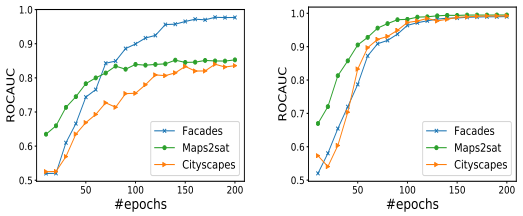
Model	Dataset	Ours	In-Dist		Out-of-Dist (BDD)	
		ROC	ROC	Acc.	ROC	Acc.
Pix2pix	Maps2sat	**85.65%**	80.15%	73.4%	-	-
Pix2pix	Cityscapes	**83.23%**	78.68%	67.5%	72.57%	56.16%
Pix2pixHD	Maps2sat	**99.42%**	98.63%	93.7%	-	-
Pix2pixHD	Cityscapes	**99.09%**	96.39%	64.0%	95.78%	56.5%

Table 9. Comparison between our MIA and the commonly used shadow-model-based classifier attack, using $4K$ train and $4K$ test images from the BDD dataset. Our MIA outperforms while not requiring extra training images.

We compare the performance of our single-sample method to the shadow model method in Tab. 9. For fairness we compare both the ROCAUC over the classifier's confidence, as well as the classification accuracy. It can be seen that in both cases, and for either in-distribution or out-of-distribution auxiliary data, the shadow model approach is inferior to our method for image translation models. We discuss the case of semantic segmentation in Sec. 4.2.2.

A.9. Memorization

As mentioned in Sec. 5, memorization is the main reason for the success of our method. Fig. 11 shows the accuracy of our method as a function of the number of epochs used for training the victim model, clearly suggesting that memorization is indeed the vulnerability.



(a) Pix2pix (b) Pix2pixHD

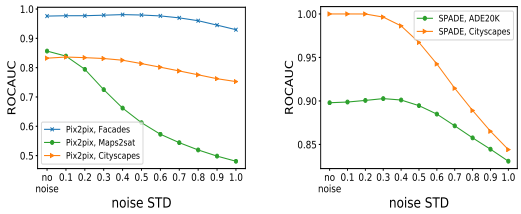
Figure 11. Effect of memorization on the attack success rate.

A.10. Defenses

In Sec. 5, we discuss the Gauss defense, including other common defenses, against our attack. We evaluated our attack accuracy as a function of different noise STD. Fig. 12 shows that a considerable amount of noise, which corrupts the generated output, is required in order to have a significant effect over our attack success, which is still much better than random guessing (50%). Results for Pix2PixHD, UperNet and HRNetV2 are presented in Fig. 5.

A.11. ImageNet predictability error

Our predictability error relies on learning a mapping between feature vectors to their corresponding pixel values. We use a pre-trained Wide-ResNet50×2 [50], which is trained



(a) Pix2pix (b) SPADE

Figure 12. Effect of Gauss defense on the attack success rate. Even with large amounts of added noise, our attack still manages to success much better then random guessing.

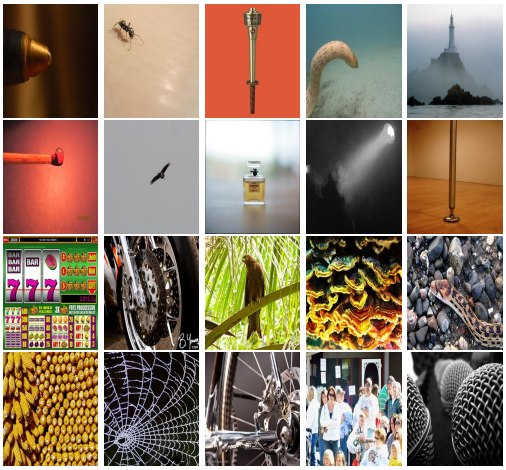


Figure 13. Examples of images from the ImageNet dataset that received the lowest and highest predictability errors. First row - lowest scored train images. Second row - lowest scored test images. Third row - highest scored train images. Last row - highest scored test images. As can be seen, the predictability error is effective even on images that were used for training the feature extractor.

on the ImageNet dataset. We do not make any assumptions regarding an overlap between the pre-trained model's training data (i.e. ImageNet) and the data during in the attack. In the scenario in which such an overlap exists, the concern is that the predictability error would lose its credibility.

In order to verify this, we computed the predictability error of a random subset of 1K train images and 1K test

images, from the ImageNet dataset. As no input-output pairs exists, we trained the linear predictor to predict pixel values from deep features of the same image. We do not observe any significant difference between the two - both share similar mean and std values: $(0.0549, 0.018)$ for the train images and $(0.0556, 0.0191)$ for the test images. A ROCAUC score of 51% further demonstrates that there is no clear difference between the distribution of the predictability error on seen and unseen images.

Fig. 13 further demonstrates this. The first row presents the train images that received the lowest scores, i.e. marked as easy images, and the second row presents the test images with the lowest scores. Both correspond to "plain" images, regardless of whether they are known (train) or unknown (test). The same applies to the Difficult images. The third row presents the highest scored train images and the last row presents the highest scored test images. Both contains complex patterns and high variance. This demonstrates that the predictability error is not affected by the having prior knowledge of the image, and is only measuring the amount of variance and complexity of an image.