One-Shot Synthesis of Images and Segmentation Masks

Vadim Sushko[1] Dan Zhang[1,2] Juergen Gall[3] Anna Khoreva[1,2]
[1]Bosch Center for Artificial Intelligence [2]University of Tübingen [3]University of Bonn
{vadim.sushko,dan.zhang2,anna.khoreva}@bosch.com, gall@iai.uni-bonn.de

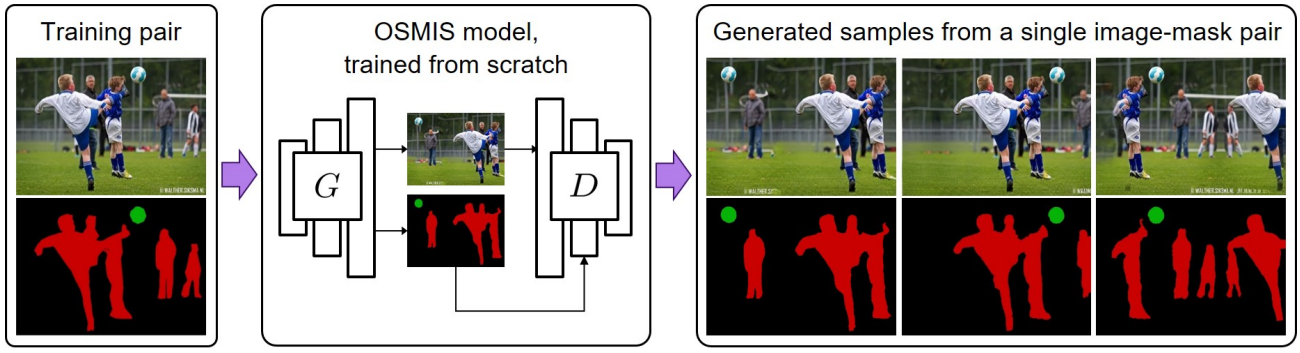

Figure 1. We introduce a new task of generating new images and their segmentation masks from a single training pair, without access to any pre-training data. Under this challenging regime, our proposed GAN model (OSMIS) achieves a synthesis of a high structural diversity, preserving the photorealism of original images and a precise alignment of produced segmentation masks to the generated content.

Abstract

Joint synthesis of images and segmentation masks with generative adversarial networks (GANs) is promising to reduce the effort needed for collecting image data with pixel-wise annotations. However, to learn high-fidelity image-mask synthesis, existing GAN approaches first need a pre-training phase requiring large amounts of image data, which limits their utilization in restricted image domains. In this work, we take a step to reduce this limitation, introducing the task of one-shot image-mask synthesis. We aim to generate diverse images and their segmentation masks given only a single labelled example, and assuming, contrary to previous models, no access to any pre-training data. To this end, inspired by the recent architectural developments of single-image GANs, we introduce our OSMIS model which enables the synthesis of segmentation masks that are precisely aligned to the generated images in the one-shot regime. Besides achieving the high fidelity of generated masks, OSMIS outperforms state-of-the-art single-image GAN models in image synthesis quality and diversity. In addition, despite not using any additional data, OSMIS demonstrates an impressive ability to serve as a source of useful data augmentation for one-shot segmentation applications, providing performance gains that are complementary to standard data augmentation techniques. Code is available at `https://github.com/boschresearch/one-shot-synthesis`.

1. Introduction

Deep neural networks have been shown powerful at solving various segmentation problems in computer vision [8, 10, 14, 23, 21, 32]. The success of these segmentation models strongly relies on the availability of a large-scale collection of labelled data for training. Nevertheless, annotation of a large dataset is not always feasible in practice due to a very high cost of manual labelling of segmentation masks [7]. For example, accurately labelling a single image with many objects can take more than 30 minutes [35]. Therefore, diminishing the human effort required for obtaining diverse and precisely aligned image-mask data is an important problem for many practical applications.

Recently, several works [30, 35, 15, 26] proposed to tackle this issue by jointly generating images and segmentation masks with generative adversarial networks (GANs). Utilizing a few provided pixel-level annotations in addition to an image dataset for training, such GAN models become a source of labelled data that can be used to train neural networks in various practical applications. Despite achieving impressive synthesis of segmentation masks based on limited annotated examples, existing image-mask GAN models still require large pre-training image datasets to learn high-fidelity image synthesis. This naturally restricts their application only to the data domains where such datasets are available (e.g., images of faces or cars). However, in some practical scenarios such a dataset can be difficult to find, for

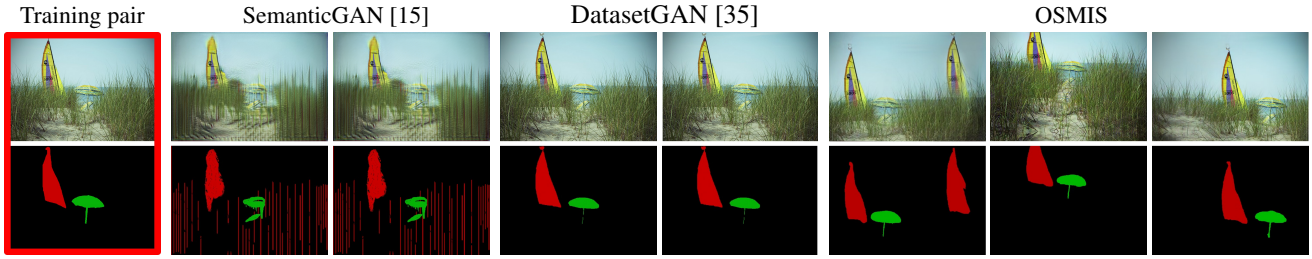

Figure 2. A comparison to SemanticGAN [15], trained on a single image-mask pair (in red), and DatasetGAN [35], pre-trained on a single image and trained on a single manual mask annotation. Both models suffer from memorization, while SemanticGAN also has poor quality due to training instabilities. In contrast, OSMIS avoids mode collapse and generates diverse high-quality samples. This is achieved by means of a discriminator that judges the realism of different objects separately, which prevents memorization of the whole given image.

example in one-shot segmentation applications [1], where the object types can be rare. Therefore, in this work we aim to learn a high-fidelity joint mask and image synthesis having as little limitations on the data domain as possible. To this end, we propose a novel GAN training setup, in which we assume availability only of a single training image and its segmentation mask, not relying on any image dataset for pre-training (see Fig. 1). After training, we aim to generate diverse new image samples and supplement them with accurate segmentation masks. To the best of our knowledge, we are the first to consider such a training scenario for GANs.

Training a GAN from a single training sample is well known to be challenging due to the problem of memorization [20], as in many cases the generator converges to reproducing the exact copies of training data. For example, as shown in our experiments, this issue occurs in the prior image-mask GAN models from [15, 35] (see Fig. 2). Recently, the issue of memorization has been mitigated in the line of works on single-image GANs, which enabled diverse image synthesis from a single training image [27, 12, 28]. Inspired by these models, we aim to extend this ability to a joint synthesis of images and segmentation masks. To this end, we propose a new model, introducing two modifications to conventional GAN architectures. Firstly, we introduce a mask synthesis branch for the generator, enabling the synthesis of segmentation masks in addition to images. Secondly, to ensure that the produced segmentation masks are precisely aligned to the generated image content, we propose a masked content attention module for the discriminator, allowing it to judge the realism of different objects separately from each other. This way, to fool the discriminator, the generator is induced to label synthesized images accurately. In effect, our proposed model enables a structurally diverse, high-quality **o**ne-**s**hot joint **m**ask and **i**mage **s**ynthesis (see Fig. 1), and we thus name it **OSMIS**. As we show in our experiments, compared to prior single-image GANs [27, 12, 28], OSMIS not only offers an additional ability to generate accurate segmentation masks, but also achieves higher quality and diversity of generated images.

Despite using only a single image-mask pair for training, OSMIS can generate a set of labelled samples of a high structural diversity, which sometimes cannot be achieved with standard data augmentation techniques (e.g., flipping, zooming, or rotation). For example, for a given scene, OSMIS can change the relative locations of foreground objects or edit the layout of backgrounds (see Fig. 1, 4, 5). Moreover, in contrast to [15, 35], OSMIS can successfully handle masks of different types, e.g., having class-wise (see Fig. 1) or instance-wise (see Fig. 4) annotations. This suggests a good potential of our model to serve as a source of additional labelled data augmentation for practical applications. We demonstrate this potential in Sec. 4.2, where we apply OSMIS at the test phase of one-shot video object segmentation [23] and one-shot semantic image segmentation [1]. The results indicate that the data generated by OSMIS helps to improve the performance of state-of-the-art networks: OSVOS [6], STM [22], and RePRI [5], providing complementary gains to standard data augmentation. We find these results promising for utilization of one-shot image-mask synthesis in future research.

2. Related Work

GANs generating segmentation masks. Recently, it was observed that a GAN generator, trained on a large dataset, implicitly learns discriminative pixel-wise features of the generated scene objects [30]. Thus, several works proposed to collect feature activations from different generator layers and transform them into a segmentation mask using a small decoder. RepurposeGAN [30] and DatasetGAN [35] proposed to train the decoder using a handful of manually annotated generated images. LinearGAN [33] replaced manual annotations by the predictions of an external segmentation network. Alternatively, SemanticGAN [15] and EditGAN [18] enforced the alignment between generated images and masks with the loss from an additional discriminator, which takes both images and masks as inputs.

Although the above models require only a few masks to achieve high-quality image-mask synthesis, they are not successful when the number of training images is not sufficient. For example, DatasetGAN and SemanticGAN suffer from instabilities and memorization issues when trained on a single image-mask pair (see Fig. 2 and A in the sup-

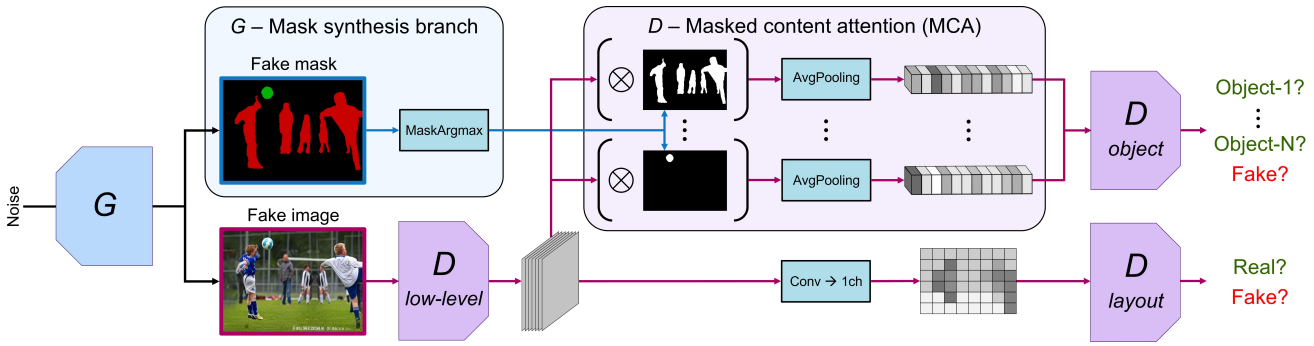

Figure 3. OSMIS model. A simple mask synthesis branch in the generator G allows the generation of segmentation masks of objects together with images. The precise alignment between the masks and the generated image content is enforced by a masked content attention (MCA) module in the discriminator D, designed to evaluate the realism of different objects separately from each other.

plementary material.). In contrast, our model learns in this regime successfully, as it does not rely on large-scale pre-training data. As shown in experiments, this makes our model better suited for the scenarios dealing with restricted data domains, such as one-shot segmentation applications. Furthermore, our model is trained in a purely adversarial fashion without any additional overhead, e.g., not requiring manual annotations of generated images, external segmentation networks, or additional discriminators.

Single Image GANs. A line of works investigated unconditional GAN training using only a single image. Under such critically low-data regime, the models are susceptible to training instabilities, as the discriminator can simply memorize the training sample and provide uninformative gradients to the generator [13]. SinGAN [27] proposed to mitigate this issue using a cascade of GANs, where each GAN stage is restricted to learn only the patch distribution at a certain image scale. ConSinGAN [12] improved the performance and efficiency of SinGAN by rebalancing the training of different GAN stages and by training several stages concurrently. Since then, numerous further variations of multi-stage GAN training have been proposed [2, 9, 4, 11]. More recently, One-Shot GAN [28] proposed a two-branch content-layout discriminator, trained as a single stage, enabling the synthesis of images with content and layouts significantly differing from the original sample. Our paper has a similar motivation to the above works, since we also aim to train a GAN model on a single data instance. However, we extend the single image setup with the synthesis of segmentation masks, which no prior work has considered, to the best of our knowledge.

3. Method

Given a single image with its pixel-level segmentation mask and assuming no access to any pre-training data, we aim to generate a diverse set of new image-mask pairs. In this section, we present OSMIS, our one-shot image-mask synthesis model. Adopting One-Shot GAN [28] as a state-of-the-art image synthesis baseline (Sec. 3.1), we propose modifications to the generator and discriminator architecture, enabling one-shot synthesis of segmentation masks that are precisely aligned with generated images (Sec. 3.2).

3.1. One-Shot GAN baseline

As the baseline network architecture, we select the state-of-the-art model One-Shot GAN [28], as it achieves the highest quality and diversity of one-shot image synthesis among previous works. One-Shot GAN proposed a two-branch discriminator, in which an input image x is first transformed into a feature representation $F(x)$ by a low-level discriminator $\mathcal{D}_{low-level}$. Next, two separate discriminators assess different aspects of $F(x)$. The content discriminator $\mathcal{D}_{content}$ judges the realism of objects regardless of their spatial location by averaging out the spatial information contained in $F(x)$ via global average pooling. On the other hand, the layout discriminator $\mathcal{D}_{layout}$ evaluates the realism only of the spatial scene layouts by squeezing $F(x)$ with a one-channel convolution. In addition, the discriminator applies feature augmentation in the content and layout representations of $F(x)$ to further increase the high-level diversity among generated samples. The adversarial loss of the One-Shot GAN model consists of three terms:

$$\mathcal{L}_{adv}(G, D) = \mathcal{L}_{D_{content}} + \mathcal{L}_{D_{layout}} + 2\mathcal{L}_{D_{low\text{-}level}}, \quad (1)$$

where each term is the mean of binary cross entropies obtained at different layers of respective discriminator parts.

3.2. OSMIS model

In contrast to one-shot image synthesis, we assume that the single training image is provided with its pixel-level mask of objects, not assuming any fixed annotation type (e.g., class-wise or instance-wise). To incorporate it into the training process, we introduce two modifications to the architecture of the baseline model. Firstly, we propose to generate segmentation masks simultaneously with images via an additional generator's mask synthesis branch. Secondly, to enforce the precise mask alignment to the generated image content, we re-formulate the objective of the content discriminator $\mathcal{D}_{content}$, designing it to judge the fidelity of different objects separately from each other. This

is made possible by the introduced masked content attention module, which builds a separate content feature vector for each object considering the provided segmentation mask. The overview of our model architecture is shown in Fig. 3. Next, we describe the proposed modifications in detail.

Mask synthesis branch in the generator. In line with [30, 35], we hypothesize that during training the generator should be able to learn discriminative features that completely describe the appearance of generated objects. Thus, while synthesizing an image, we collect feature activations of the generator layers and use them as input for the mask synthesis branch. In contrast to [30, 35], we use only the activations after the last generator block, as this simplest solution already performs well in our experiments. Using a simple convolution followed by a softmax activation, we transform these features into an N-channel soft probability map, where each channel corresponds to one of $N-1$ objects of interest in the segmentation mask or to the background. To obtain the final discrete mask prediction, an argmax operation T along the channel dimension is applied.

To enable the training of the mask synthesis branch with the discriminator loss, the generated masks should allow back-propagation of gradients, similarly to generated images. In our experiments, feeding the discriminator the continuous segmentation probability maps obtained before the non-differentiable argmax operation T impaired the GAN training, as the discriminator learnt to detect the continuous-discrete discrepancy between fake and real inputs. Thus, inspired from [31, 3], we enable back-propagation through argmax by developing a straight-through gradient estimator:

$$\text{MaskArgmax}(y) = y + T(y) - sg[y], \tag{2}$$

where sg denotes a stop-gradient operation. This way, the discriminator is provided with the generated masks in a discrete form $T(y)$, which enables its effective training, while the generator can be trained with the gradients passing through its probability map prediction y.

Yet, this solution can sometimes lead to degenerate solutions, e.g., when all the pixels are predicted as the background channel. This cannot be corrected during training, as in this case the gradient flow through all the other mask channels is blocked. We found that it can be mitigated by softening the argmax operation T at the beginning of training. For this, during the first P_0 epochs we regard each mask pixel as a random variable following Bernoulli distribution:

$$T(y) = \begin{cases} \sim \text{Bernoulli}(y) & \text{epoch} < P_0, \\ \text{argmax}(y) & \text{epoch} \geq P_0. \end{cases} \tag{3}$$

Masked content attention in the discriminator. To provide a training signal to the generator's mask synthesis branch, we propose to incorporate the learning of the image-mask alignment to the objective of the content discriminator $\mathcal{D}_{content}$. In [28], $\mathcal{D}_{content}$ was designed to judge the content distribution of the whole given image. Considering the provided segmentation mask, we can now select the image areas belonging to different objects, and require the discriminator to learn their appearance separately from each other. With this objective, as the discriminator can compare the appearance of the area belonging to the same object in real and fake images, it encourages the generator not only to synthesize realistic objects, but also to label them correctly.

To this end, we introduce a masked content attention (MCA) module. As shown in Fig. 3, MCA receives a downsampled segmentation mask y along with an intermediate feature representation $F(x) = \mathcal{D}_{low-level}(x)$ of an input image x, and thereout produces a set of N content vectors, corresponding to the masked content representations of each of the $N-1$ objects of interest and the background:

$$\text{MCA}(x, y) = \{\text{AvgPool}\left(F(x) \times \mathbb{1}_{y=i}\right)\}_{i=1}^{N}. \tag{4}$$

Accordingly, we re-design the objective of the content discriminator (further denoted $\mathcal{D}_{object}$). For each of the obtained object representations, our proposed $\mathcal{D}_{object}$ is induced to predict a correct identity of each object or background of a real image, while all the identities of fake images should be categorized as an additional fake class:

$$\begin{aligned} \mathcal{L}_{D_{object}} = &- \mathbb{E}_{(x,y)}\left[\sum_{i=1}^{N} \alpha_i \log \mathcal{D}^i_{object}(\text{MCA}^i(x, y))\right] \\ &- \mathbb{E}_z\left[\sum_{i=1}^{N} \log(1 - \mathcal{D}^{fake}_{object}(\text{MCA}^i(G(z)))\right], \end{aligned} \tag{5}$$

where z is the noise vector used by the generator G to synthesize a fake image-mask pair $G(z) = \{G_x(z), G_y(z)\}$, (x, y) denotes the real image-mask pair, and $\mathcal{D}^i(*)$ is the discriminator logit for the object i. Considering that different objects or background can occupy different areas, we introduce a class balancing weight α_i, which is the inverse of the per-pixel class frequency in the segmentation mask y:

$$\alpha_i = \frac{(\text{sum}(\mathbb{1}_{y=i}))^{-1}}{\sum_{j=1}^{N}(\text{sum}(\mathbb{1}_{y=j}))^{-1}}. \tag{6}$$

Note that the balancing is applied only for real images, as in Eq. 5 all fake objects are considered as the same class.

Our $\mathcal{D}_{object}$ learns the content distribution of each object separately. The advantage of such a training scheme is two-fold. Firstly, a generator now needs to synthesize correct segmentation masks in order to fool the discriminator. The precise image-mask alignment is thus enforced directly by the adversarial loss, without the need for using additional networks or manual annotation. Secondly, as MCA provides representations only of separate objects, $\mathcal{D}_{object}$ has restricted access to the content distribution of the whole image. In effect, the discriminator memorization of the whole training sample becomes more difficult, which enables more diverse image synthesis (see Table 3).

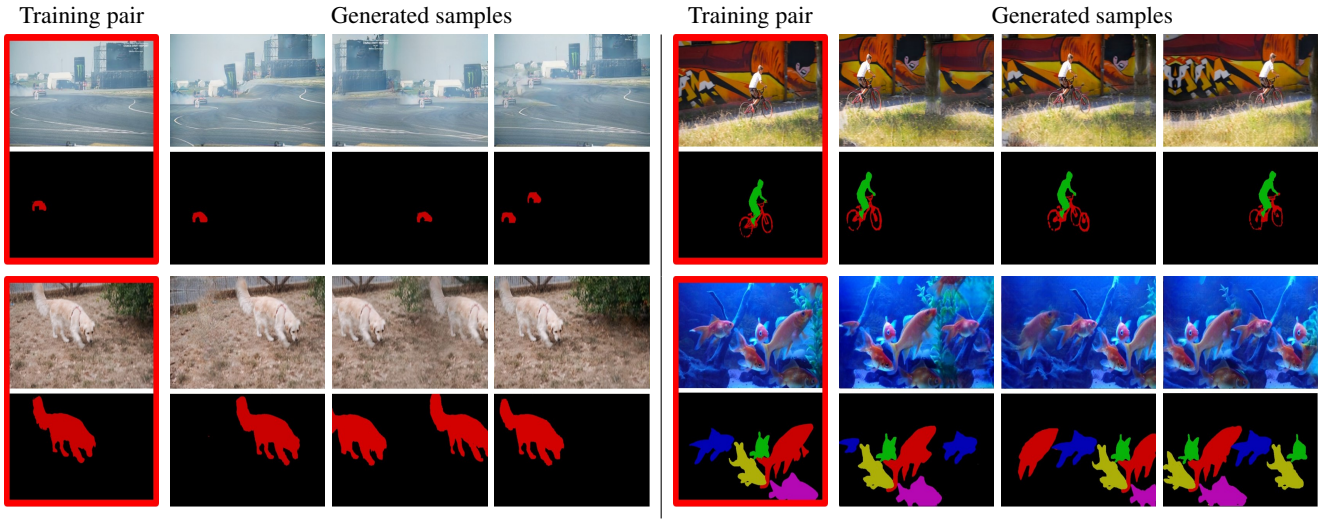

Figure 4. Qualitative results of OSMIS on DAVIS [23]. Given a single image-mask pair for training, our model achieves high-fidelity image synthesis with a high structural diversity, changing the positions of objects or editing the layout of backgrounds. For each synthesized image, it produces segmentation masks that accurately annotate the generated content. Training pairs are shown in red frames.

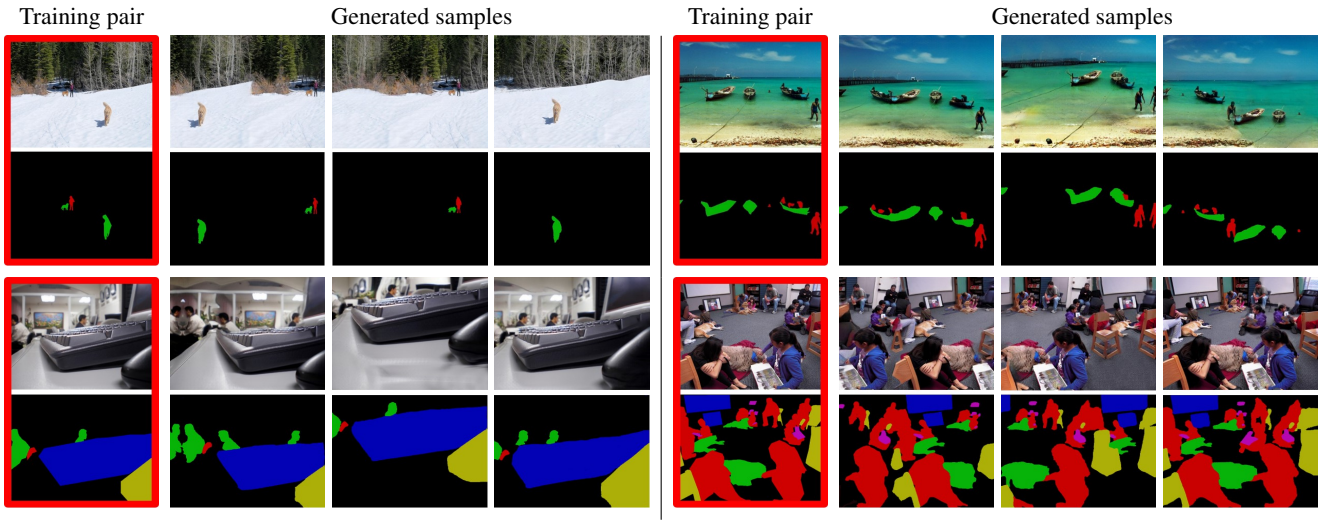

Figure 5. Qualitative results of OSMIS on COCO [17]. OSMIS successfully deals with different scene types and annotation styles. For example, it achieves high quality and diversity for both indoor and outdoor scenes, or sparse and dense annotations of foreground objects.

4. Experiments

We evaluate our model as follows. Firstly, we provide the qualitative and quantitative assessment of the achieved one-shot image-mask synthesis, evaluating the quality and diversity of generated images, as well as their alignment to the produced segmentation masks (Sec. 4.1). Secondly, we apply OSMIS to two one-shot segmentation applications, demonstrating the potential of the generated image-mask pairs to be used as data augmentation (Sec. 4.2).

4.1. Evaluation of one-shot image-mask synthesis

Training details. We train our model with the loss from Eq. (5) for the object discriminator $\mathcal{D}_{object}$, setting P_0=15000. We employ differentiable augmentation (DA) of input images and masks while training the discriminator, using the whole set of transformations as proposed in [13]. We use an exponential moving average of the generator weights with a decay of 0.9999, and follow [28] in setting all the other hyperparameters. More training details are shown in the supplementary material.

Datasets. To evaluate the synthesis, we use the DAVIS dataset [23], originally introduced for video object segmentation. For each video from the DAVIS-17 validation split, we take the first frame and its segmentation mask of objects, which results in 30 image-mask pairs on which we train separate models. The resolution is set to 640x384. For additional visual results, we use samples from COCO [17], trying to closely fit their resolution. Note that the datasets have

Method	SIFID↓	LPIPS↑
SinGAN [27]	0.131	0.267
ConSinGAN [12]	0.103	0.296
One-Shot GAN [29]	0.091	0.347
OSMIS (ours)	**0.073**	**0.387**

Table 1. Comparison of image quality and diversity to single-image GANs on DAVIS-17. Bold denotes the best performance.

Method	SIFID↓	LPIPS↑	mIoU
DatasetGAN [35]	0.118	0.007	91.1*
SemanticGAN [15]	0.211	0.012	65.8
OSMIS (ours)	**0.073**	**0.387**	**86.6**

Table 2. Comparison to prior image-mask GANs on DAVIS-17. Bold denotes the best performance. Red indicates mode collapse. * Indicates manual annotation of masks for DatasetGAN training.

different annotation types (class-wise and instance-wise).

Metrics. To mind a possible quality-diversity trade-off in our one-shot regime [24, 16], we assess the quality and diversity of generated images separately. For this, we report the average SIFID [27] as the measure of image quality, while the average LPIPS [34] between the pairs of generated images is used to assess the diversity of synthesis.

On the other hand, evaluating the quality of generated masks is challenging, because generated images do not have ground truth segmentation annotations. To bypass this issue, we propose to evaluate the alignment between generated masks and synthetic images using an external segmentation network. For this, we take a UNet [25] and train it on the generated image-mask pairs for 500 epochs. After training, we compute its mIoU performance on the original real image, augmented with standard geometric transformations. Intuitively, a good performance on this test reveals that synthetic masks describe well the objects from the real data, indicating precise alignment between the generated images and their masks.

Qualitative results. Fig. 4 and 5 show image-mask pairs generated by OSMIS trained on samples from DAVIS and COCO. Given only a single image-mask pair, our model learns to generate new image-mask pairs, demonstrating a remarkable structural diversity among samples, photorealism of synthesized images, and a high quality of generated annotations. For example, OSMIS can re-synthesize the provided scene with a different number of foreground objects, e.g., more dogs (3rd example in Fig. 4), less people (2nd example in Fig. 5), or edit layouts of backgrounds (1st examples in Fig. 4-5), in all cases providing accurate segmentation masks for the re-synthesized scenes. We note that reaching such structural differences to training data simultaneously with photorealism is extremely difficult from a single sample. For example, it could not be achieved with DatasetGAN or SemanticGAN due to memorization issues and training instabilities (see Fig. 2). Lastly, we remark that

Mask supervision	SIFID↓	LPIPS↑	mIoU
None	0.071	0.368	-
Projection [19]	**0.071**	0.362	72.1
Input concat.	0.079	0.328	82.4
SemanticGAN D_m [15]	0.074	0.351	83.3
MCA (ours)	0.073	**0.387**	**86.6**

Table 3. Comparison of MCA to other mask synthesis supervision mechanisms on DAVIS-17. Red indicates decreased diversity compared to the baseline. Bold denotes the best performance.

OSMIS successfully deals with very different scene types (e.g., both indoor and outdoor scenes), supports masks with both sparse and dense object annotations (e.g., foreground objects occupying small or large image areas), and can handle masks with many objects or even separate instances of the same semantic class (e.g., fish in 4th example in Fig. 4).

Quantitative results. We compare the quality and diversity of generated *images* to the single-image GAN models SinGAN [27], ConSinGAN [12] and One-Shot GAN [28]. The image-mask synthesis is compared to the previous methods DatasetGAN [35] and SemanticGAN [15]. We use the official repositories provided by the authors.

The quantitative comparison of the image synthesis to single-image GAN models on DAVIS-17 is presented in Table 1. Compared to these models, OSMIS not only offers an additional ability to generate segmentation masks, but also achieves higher image quality and diversity. As seen in Table 1, despite a potential trade-off between SIFID and LPIPS, our model outperforms previously published baselines in both metrics by a notable margin. Further, Table 2 demonstrates that prior image-mask methods, DatasetGAN and SemanticGAN, suffer from instabilities and fail to achieve diverse synthesis, scoring very low in LPIPS.

Ablations. In Table 3 we compare the proposed masked content attention module (MCA) with three alternative discriminator mechanisms to provide supervision for the generator's mask synthesis branch. The simplest baseline is to concatenate the input masks to images, requiring the discriminator to judge their realism jointly. Another method is to use projection [19], by taking the inner product between the last linear layer output of $D_{\text{low-level}}$ and the pixel-wise linear projection of the input mask. Finally, we compare to the approach of SemanticGAN [15], adding a separate discriminator network $\mathcal{D}_m$ which takes both segmentation masks and images, and propagate its gradients only to the generator's mask synthesis branch. While training these baselines, we preserve all the OSMIS hyperparameters, but remove the MCA and use the original $\mathcal{D}_{content}$ as in [28]. As seen from mIoU in Table 3, MCA enables the generation of segmentation masks with the best alignment to the generated image content, as measured by an external segmentation network. Notably, while all the alternative methods negatively affect diversity, MCA improves it (0.387 vs 0.368 LPIPS), highlighting its regularization effect which

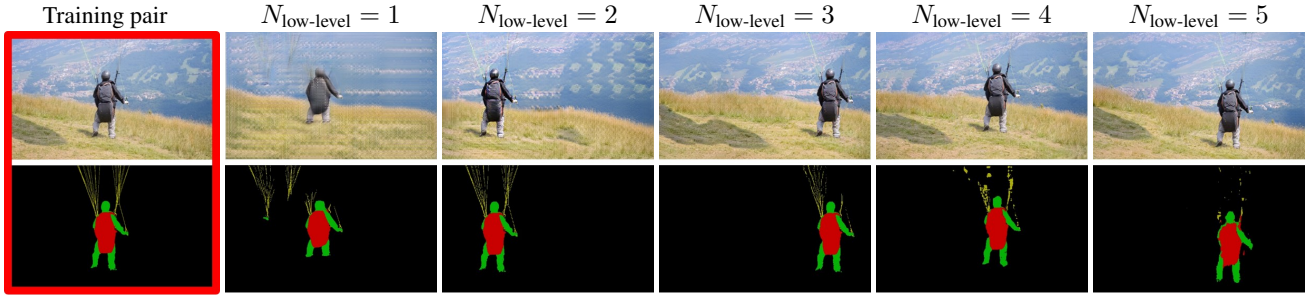

Figure 6. Trade-off between the image and mask quality when varying the number of $\mathcal{D}_{\text{low-level}}$ discriminator blocks. Increased number improves image quality, but harms the ability of masks to capture fine-grained object details due to stronger downsampling during training.

$N_{\text{low-level}}$	SIFID↓	LPIPS↑	mIoU
1	0.262	0.395	82.4
2	0.165	**0.404**	**87.1**
3	0.102	0.394	86.9
4	0.073	0.387	86.6
5	**0.070**	0.321	83.9

Table 4. Ablation on the number of $\mathcal{D}_{\text{low-level}}$ discriminator blocks on DAVIS-17. Bold denotes the best performance.

prevents the discriminator memorization of training data.

While enabling on average higher image diversity and mask quality, we found that MCA can struggle if the training sample contains annotations of fine-grained object details, due to downsampling of input masks. This is illustrated in Fig. 6 and Table 4, for which we train OSMIS with different numbers of low-level discriminator blocks $N_{\text{low-level}}$, corresponding to different degrees of mask downsampling. We observe a trade-off between the quality of images and masks: decreasing $N_{\text{low-level}}$ improves the image diversity and pixel-level mask fidelity, but harms image quality. We selected $N_{\text{low-level}} = 4$ as a compromise between the metrics in Table 4, even though this configuration sometimes fails to annotate small object details (as in Fig. 6). Note that despite this limitation, MCA still outperforms alternative methods that do not use downsampling on DAVIS-17 (see Table 3), and leads to image-mask pairs that are more useful as data augmentation, as discussed next.

4.2. Application to one-shot segmentation tasks

After training, OSMIS can augment the provided image-mask pair with novel diverse samples. As such diversity (edited backgrounds, objects changing relative locations) is difficult to achieve by means of standard data augmentation, we foresee a potential usage of our model as a source of labelled data augmentation. Thus, in what follows, we test the efficacy of OSMIS generations when applied at test phase of two one-shot segmentation applications.

One-shot video object segmentation. We apply our model to the semi-supervised one-shot video segmentation benchmark DAVIS [23]. At test phase, this task provides a video and the segmentation mask of objects only in the first frame, while a model is required to segment all the remaining video frames. We select two popular models from the literature: OSVOS [6], which fine-tunes the network weights on the first video frame and segments other frames independently, and STM [22], which propagates the segmentation prediction sequentially using a space-time memory module. We conduct experiments on two DAVIS splits: *DAVIS-16*, having 20 videos with a single annotated object; and its extension *DAVIS-17*, having 30 videos with multi-instance annotations. To evaluate the video segmentation, we compute the average of the mean mIoU region similarity ($\mathcal{J}$) and the mean contour accuracy ($\mathcal{F}$) across all videos, which is a popular metric for this task [23].

Network	Augmentation: Standard	Augmentation: **Ours**	DAVIS-16	DAVIS-17
OSVOS [6]	✗	✗	76.9	51.3
	✓	✗	78.5 (80.2)	52.9 (52.8)
	✗	✓	78.2	52.6
	✓	✓	**79.8**	**54.2**
STM [22]	✗	✗	89.7 (89.4)	72.4 (72.2)
	✓	✗	89.9	72.4
	✗	✓	90.1	72.6
	✓	✓	**90.2**	**72.7**

Table 5. Effect of data augmentation on the mean of mIoU and contour accuracy ($\mathcal{J}\&\mathcal{F}$) of one-shot video object segmentation. Bold denotes the best performance. Round brackets show the results reported in [6, 22]. Reproduced and reported numbers for OSVOS differ as its official code lacks some model components.

One-shot semantic image segmentation. The second setup is the one-shot image segmentation benchmark COCO-20^i [17]. In this task, a segmentation model is first trained on a large dataset. At test phase, the model is given a single image-mask pair (support set) with an object of a previously unseen test class, and is then required to segment another sample (query image) containing instances of the same class. We conduct experiments with the state-of-the-art RePRI network [5]. COCO-20^i contains 80 classes, which are divided into 4 folds, with 60 base and 20 test classes in each fold. To test OSMIS, we randomly selected 5 support samples for each test class, resulting in 100 image-

Network	Augmentation: Standard	Augmentation: **Ours**	$COCO^0$	$COCO^1$	$COCO^2$	$COCO^3$
RePRI [5]	✗	✗	31.2 (31.2)	38.3 (38.1)	32.9 (33.3)	33.2 (33.0)
	✓	✗	31.8	38.5	33.4	33.8
	✗	✓	32.4	38.7	33.7	34.3
	✓	✓	**32.8**	**39.0**	**34.1**	**34.6**

Table 6. Effect of synthesized data augmentation on mIoU of one-shot image segmentation. In each data split, support examples were sampled from a subset of 100 image-mask pairs, for which our model was trained. Bold denotes the best performance. The round brackets contain the numbers reported in [5].

mask pairs in each of the folds, and trained OSMIS on all of them separately. The performance of this task is evaluated separately for each fold, using the average mIoU across many different support-query examples.

Experimental setup. For both applications, we train OSMIS on the single given image-mask pair (the first video frame or support sample). We try to closely fit the resolution of each image from COCO, and set a fixed resolution of 640x384 for images from the DAVIS benchmark. After training, we generate a pool of synthetic image-mask pairs consisting of $n = 100$ samples. As OSMIS can occasionally fail and synthesize noisy examples, we compute the SIFID metric [27] for each generated image as a measure of its quality. Ranking the images by the average of SIFID ranks at different InceptionV3 layers, we exclude bad-quality samples by filtering out 15% lowest-ranked images. Finally, we add the remaining synthetic samples to the original image-mask pair as data augmentation. See more setup details in Sec. B of the supplementary material.

Among the used segmentation models, only OSVOS [6] applies data augmentation at test phase (random combinations of image-mask flipping, zooming, and rotation). Thus, in experiments we compare our synthetic data augmentation to this pipeline (referred to as *standard* augmentation).

Results. The performance of segmentation networks using different data augmentation is shown in Tables 5 and 6. To account for the variance between runs, all the results are averaged across 5 runs with different seeds for augmentation. We generally managed to reproduce the official reported numbers closely, with the exception of OSVOS, for which the official codebase[1] does not implement the model in full configuration. As seen in Tables 5 and 6, the synthetic data augmentation produced by OSMIS yields a notable increase in segmentation performance, on average improving the metrics of OSVOS and STM by 1.3 and 0.3 $\mathcal{J}\&\mathcal{F}$ points, and RePRI by 0.9 mIoU points compared to the models using no data augmentation. Despite a possible mismatch between OSMIS training resolution and target image size (e.g., 640x384 vs 854x480 for DAVIS) and the need for image resizing, our synthetic data augmentation consistently outperforms standard data augmentation for STM and RePRI, and is almost on par for OSVOS, which was originally tuned for training with standard data augmentation. These results validate the ability of OSMIS to generate structurally diverse data augmentation of sufficient quality in the one-shot regime. Finally, we note that the effect of OSMIS generations is complementary to standard data augmentation, as the best results for all models are observed when the two pipelines are used in combination.

[1] https://github.com/kmaninis/OSVOS-PyTorch

Synthesis method	OSVOS, DAVIS-16 $\mathcal{J}\&\mathcal{F}$	RePRI, $COCO^0$ mIoU
Reference w/o synth. augm.	78.5	31.8
SemanticGAN [15]	73.1	29.4
DatasetGAN [35]	77.8	30.9
Projection [19]	78.4	30.9
Input concat.	79.3	31.9
SemanticGAN D_m [15]	79.5	32.3
MCA (ours)	**79.8**	**32.8**

Table 7. Impact on the performance of synthesized data produced with different models and mask supervision methods. The reference performance is obtained using standard data augmentation. Bold denotes the best performance.

Table 7 demonstrates the efficiency of synthetic data augmentation obtained with different GAN models. The previous image-mask models DatasetGAN and SemanticGAN both show poor applicability in the scenario of one-shot applications due to poor synthesis performance. Further, among the comparison methods for mask synthesis supervision, the strongest increase in performance is achieved with our proposed MCA module. This indicates that the high synthesis diversity and precise image-mask alignment (see Table 3) are the keys to achieve useful data augmentation.

5. Conclusion

We presented OSMIS, an unconditional GAN model that can learn to generate new high-quality image-mask pairs from a single training pair, not relying on any pre-training data. In such a low-data regime, our model generates photorealistic scenes that structurally differ from the original samples, while the produced masks are precisely aligned to the generated image content. Although the synthesis of OSMIS is inherently constrained by the appearance of objects in the original sample, it can serve as a source of useful data augmentation for one-shot segmentation applications, providing complementary gains to standard image augmentation. Thus, we find using one-shot image-mask synthesis in practical applications promising for future research.

Aknowledgement. Juergen Gall was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2070 -390732324 and the ERC Consolidator Grant FORHUE (101044724).

References

[1] Zhen Liu Irfan Essa Amirreza Shaban, Shray Bansal and Byron Boots. One-shot learning for semantic segmentation. In *British Machine Vision Conference (BMVC)*, 2017.

[2] Rajat Arora and Yong Jae Lee. SinGAN-GIF: Learning a generative video model from a single GIF. In *Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[3] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv:1308.3432*, 2013.

[4] Raphael Bensadoun, Shir Gur, Tomer Galanti, and Lior Wolf. Meta internal learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[5] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision (ECCV)*, 2018.

[9] Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical Patch VAE-GAN: Generating diverse videos from a single sample. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Xiaoyu He and Zhenyong Fu. Recurrent SinGAN: Towards scale-agnostic single image GANs. In *International Conference on Electronic Information Technology and Computer Engineering*, 2021.

[12] Tobias Hinz, Matthew Fisher, Oliver Wang, and Stefan Wermter. Improved techniques for training single-image GANs. In *Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[13] Tero Karras, Miika Aittala, Janne Hellsten, S. Laine, J. Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[14] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[16] Yijun Li, Richard Zhang, Jingwan Cynthia Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[17] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[18] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[19] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations (ICLR)*, 2018.

[20] Vaishnavh Nagarajan, Colin Raffel, and Ian J Goodfellow. Theoretical insights into memorization in gans. In *Advances in Neural Information Processing Systems (NeurIPS) Workshops*, 2018.

[21] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[22] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *International Conference on Computer Vision (ICCV)*, 2019.

[23] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv:2010.11943*, 2021.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[26] Oindrila Saha, Zezhou Cheng, and Subhransu Maji. GANORCON: Are generative models useful for few-shot segmentation? *arXiv:2112.00854*, 2021.

[27] Tamar Rott Shaham, Tali Dekel, and T. Michaeli. SinGAN: Learning a generative model from a single natural image. In *International Conference on Computer Vision (ICCV)*, 2019.

[28] Vadim Sushko, Juergen Gall, and Anna Khoreva. One-Shot GAN: Learning to generate samples from single images and videos. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

[29] Vadim Sushko, Dan Zhang, Juergen Gall, and Anna Khoreva. Learning to generate novel scene compositions from single images and videos. *arXiv:2103.13389*, 2021.

[30] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing GANs for one-shot semantic part segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[31] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[32] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *International Conference on Computer Vision (ICCV)*, 2019.

[33] Jianjin Xu and Changxi Zheng. Linear semantics in generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[35] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. DatasetGAN: Efficient labeled data factory with minimal human effort. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.