ONE-SHOT GAN: LEARNING TO GENERATE SAMPLES FROM SINGLE IMAGES AND VIDEOS

Vadim Sushko[1], **Jürgen Gall**[2], **Anna Khoreva**[1]
[1]Bosch Center for Artificial Intelligence, [2]University of Bonn

ABSTRACT

Training GANs in extremely low-data regimes remains a challenge, mainly due to overfitting leading to memorization or training divergence. In this work, we introduce One-Shot GAN, an unconditional generative model that can learn to generate samples from a training set as little as one image or one video. We propose a two-branch discriminator architecture, with content and layout branches designed to judge the internal content separately from the scene layout realism. This encourages to generate images with varying content and global layouts while preserving the context of the original sample. Compared to previous single-image GAN models, One-Shot GAN achieves higher diversity and quality of image generation, while also not being restricted to a single-image training. We show that our model successfully deals with other one-shot regimes, and introduce a novel benchmark of learning generative models from frames of a single video.

Figure 1: Our proposed One-Shot GAN needs only one video (first two rows) or one image (last two rows) for training. From a single video with a car on a road, One-Shot GAN can generate the scene without the car or with two cars, and for a single air balloon image, it produces layouts with different number and placement of the balloons. (Original samples are shown in grey or red frames.)

1 INTRODUCTION

Without sufficient training data, GANs are prone to overfitting, which often leads to mode collapse and training instabilities (Shaham et al., 2019; Hinz et al., 2020). This dependency on availability of training data limits the applicability of GANs in domains where collecting a large dataset is not feasible. In some real-world applications, collection even of a small dataset remains challenging. It may happen that rare objects or events are present only on one image or on one video, and it is difficult to obtain a second one. This, for example, includes pictures of exclusive artworks or videos of traffic accidents recorded in extreme conditions. Enabling learning of GANs in such *one-shot* scenarios is thus an important task, having a potential to improve their utilization in practice. Previous work (Shaham et al., 2019; Hinz et al., 2020) studied one-shot image generation in the context of learning from a *single image*. In this work, we introduce a novel setup of learning to

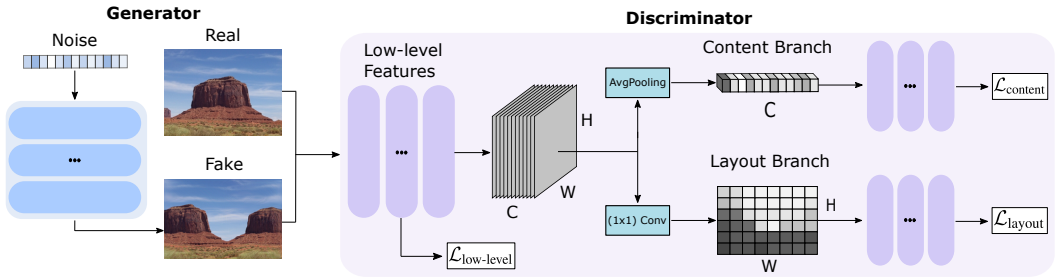

Figure 2: One-Shot GAN. The two-branch discriminator judges the content separately from the scene layout realism, enabling the generator to produce images with varying content and layouts.

generate images from frames of a *single video*. In practice, recoding a video lasting for several seconds can take almost as little effort as collecting one image. However, in comparison to a single image, video contains much more information about the scene and the objects of interest (e.g., different poses and locations of objects in the scene, various camera views, dynamic backgrounds). Learning from a video can enable generation of images of higher quality and diversity, while still operating in a one-shot mode, and therefore can improve its usability for practical applications.

Training from a short video clip requires a model to be able to learn meaningful semantic properties of frames, without simply memorizing them. As we show in our experiments, previously proposed GAN models (Shaham et al., 2019; Liu et al., 2021), designed for low-data regimes, do not succeed in this setting. To this end, we design a One-Shot GAN model, which can successfully operate in different one-shot settings, needing only one image or one short video clip for training. This is achieved by two key ingredients: the novel design of the discriminator and the proposed diversity regularization technique for the generator. The new One-Shot GAN discriminator has two branches, responsible for judging the content distribution and the scene layout realism of images separately from each other. Disentangling the discriminator decision about the content and layout helps to prevent overfitting and provides more informative signal to the generator. To achieve high diversity of generated samples, we also extend the regularization technique of (Yang et al., 2019; Choi et al.) to one-shot unconditional image synthesis. The proposed One-Shot GAN mitigates overfitting and generates high-quality images that are substantially different from training data. One-Shot GAN is the first model that succeeds in learning from both single images and videos, improving over prior work (Shaham et al., 2019; Hinz et al., 2020; Liu et al., 2021) in image quality and diversity.

2 One-Shot GAN

Content-Layout Discriminator. We introduce a solution to overcome the memorization effect but still to generate images of high-level diversity in the one-shot setting. Building on the assumption that to produce realistic and diverse images the generator should learn the appearance of objects and combine them in a globally-coherent way in an image, we propose a discriminator that judges the *content* distribution of an image separately from its *layout* realism. To achieve the disentanglement, we design a two-branch discriminator architecture, with separate content and layout branches (see Fig.2). Our discriminator D consists of the low-level feature extractor $D_{low\text{-}level}$, the content branch $D_{content}$, and the layout branch D_{layout}. Note that the branching happens after an intermediate layer in order to learn a relevant representation. $D_{content}$ judges the content of this representation irrespective from its spatial layout, while D_{layout}, contrarily, inspects only the spatial information. Inspired by the attention modules of Park et al. (2018); Woo et al. (2018), we extract the *content* from intermediate representations by aggregating spatial information via global average pooling, and obtain *layout* by aggregating channels via a simple (1×1) convolution. This way, the content branch judges the fidelity of "objects" composing the image independent of their spatial location, while the layout branch is sensitive only to the realism of global scene layouts. Note that $D_{content}$ and D_{layout} receive only limited information from previous layers, which prevents overfitting. This helps to overcome the memorization of training data and to produce different images.

Diversity regularization. To improve variability of generated samples, we propose to add diversity regularization (DR) loss term $\mathcal{L}_{DR}$ to the objective. Prior diversity regularization terms (Yang et al., 2019; Choi et al.) aimed to encourage the generator to produce different outputs depending on the input latent code, in such way that the generated samples with closer latent codes should look more similar to each other, and vice versa. In contrast, in the one-shot image synthesis setting,

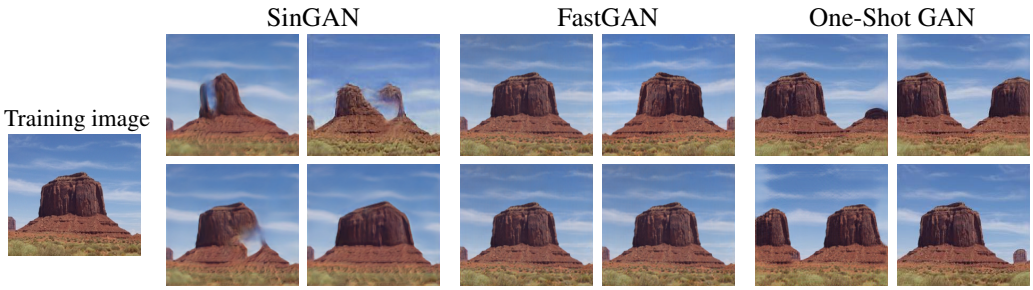

Figure 3: Comparison in the Single Image setting. SinGAN may incoherently shuffle patches (e.g. sky textures below horizon), and FastGAN can only reproduce the image or its flipped version. Contrary, One-Shot GAN achieves high diversity, preserving content distribution and realistic layouts.

Method	Single Image				Single Video			
	SIFID↓	LPIPS↑	MS-SSIM ↓	Dist. to train	SIFID↓	LPIPS↑	MS-SSIM ↓	Dist. to train
SinGAN	0.13	0.26	0.69	0.30	2.47	0.32	0.65	0.51
FastGAN	0.13	0.18	0.77	0.11	0.79	**0.43**	0.55	0.13
One-Shot GAN	**0.08**	**0.33**	**0.63**	0.37	**0.55**	**0.43**	**0.54**	0.34

Table 1: Comparison in the Single Image and Single Video settings on DAVIS-YFCC100M dataset.

the perceptual distance of the generated images should not be dependent on the distance between their latent codes. As we operate in one semantic domain, the generator should produce images that are in-domain but more or less equally different from each other and substantially diverse form the original sample. Thus, we propose to encourage the generator to produce perceptually different image samples independent of their distance in the latent space. $\mathcal{L}_{DR}$ is expressed as follows:

$$\mathcal{L}_{DR}(G) = \mathbb{E}_{z_1,z_2}\big[\frac{1}{L}\sum_{l=1}^{L}\|G^l(z_1) - G^l(z_2)\|\big], \tag{1}$$

where $\|\cdot\|$ denotes the $L1$ norm, $G^l(z)$ indicates features extracted from the l-th block of the generator G given the latent code z. Contrary to prior work, we compute the distance between samples in the feature space, enabling more meaningful diversity of the generated images, as different generator layers capture various image semantics, inducing both high- and low-level diversity.

Final objective. We compute adversarial loss for each discriminator part: $D_{low\text{-}level}$, $D_{content}$, and D_{layout}. This way, the discriminator decision is based on low-level details of images, such as textures, and high-scale properties, such as content and layout. The overall adversarial loss is

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

where $\mathcal{L}_{D_*}$ is binary cross-entropy $\mathbb{E}_x[\log D_*(x)] + \mathbb{E}_z[\log(1 - D_*(G(z)))]$ for real image x and generated image $G(z)$. As the two branches of the discriminator operate at high-level image features, contrary to only one $D_{low\text{-}level}$ operating at low-level features, we double the weighting for the $\mathcal{L}_{D_{low\text{-}level}}$ loss term. This is done in order to properly balance the contributions of different feature scales and encourage the generation of images with good low-level details, coherent contents and layouts. The overall One-Shot GAN objective can be written as:

$$\min_G \max_D \; \mathcal{L}_{adv}(G, D) - \lambda\mathcal{L}_{DR}(G), \tag{3}$$

where λ controls the strength of the diversity regularization and $\mathcal{L}_{adv}$ is the adversarial loss in Eq. 2.

Implementation. The One-Shot GAN discriminator uses ResNet blocks, following Brock et al. (2019). We use three ResNet blocks before branching and four blocks for the content and layout branches. We employ standard image augmentation strategies for the discriminator training, following Karras et al. (2020a). λ for $\mathcal{L}_{DR}$ in Eq. 3 is set to 0.15. We use the ADAM optimizer with $(\beta_1, \beta_2) = (0.5, 0.999)$, the batch size of 5 and the learning rate of 0.0002 for both G and D.

3 Experiments

Evaluation settings. We evaluate One-Shot GAN on two different one-shot settings: training on a single image and a single video. We select 15 videos from DAVIS (Pont-Tuset et al., 2017) and YFCC100M (Thomee et al., 2016) datasets. In the Single Video setting, we use all frames of a video as training images, while for the Single Image setup we use only one middle frame. The chosen videos last for 2-10 seconds and consist of 60-100 frames. To assess the quality of generated

images, we measure single FID (SIFID) (Shaham et al., 2019). Image diversity is assessed by the average LPIPS (Dosovitskiy & Brox, 2016) and MS-SSIM (Wang et al., 2003) across pairs of generated images. To verify that the models do not simply reproduce the training set, we report average LPIPS to the nearest image in the training set, augmented the same way as during training (Dist. to train). We compare our model with a single image method SinGAN (Shaham et al., 2019) and with a recent model on few-shot image synthesis, FastGAN (Liu et al., 2021).

Main results. Table 1 presents quantitative comparison between the models in the Single Image and Video settings, while the respective visual results are shown in Fig. 3 and 4. As seen from Table 1, One-Shot GAN notably outperforms other models in both quality and diversity metrics. Importantly, our model is the only one which successfully learns from both single images and single videos.

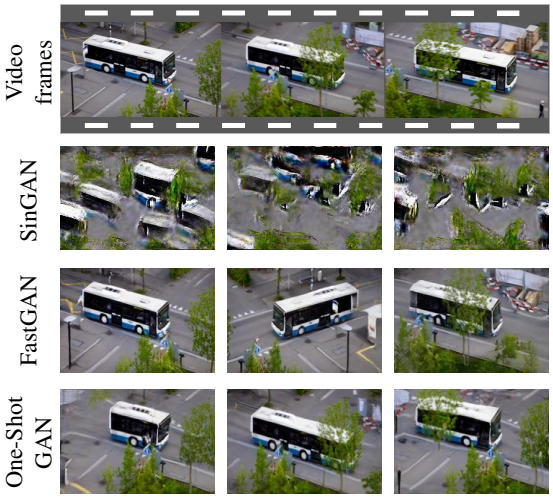

Figure 4: Comparison with other methods in the Single Video setting. While FastGAN reproduces the training frames and SinGAN fails to draw objects, One-Shot GAN generates high-quality images substantially different from original frames.

As seen from Fig. 1 and 3, in the Single Image setting, One-Shot GAN produces diverse samples of high visual quality. For example, our model can change the number of rocks on the background or change their shapes. Note that such changes keep appearance of objects, preserving *content* distribution, and maintain global *layout* realism. In contrast, single-image method SinGAN disrespects layouts (e.g. sky textures appear below horizon), and is prone to modest diversity, especially around image corners. This is reflected in higher SIFID and lower diversity scores in Table 1. Concurrently, the few-shot FastGAN suffers from memorization, only reproducing the training image or its flipped version. In Table 1 this is reflected in low diversity and small Dist. to train (in red) metrics.

Fig. 4 and 1 show images generated in the Single Video setting. One-Shot GAN produces high-quality images that are substantially different from the training frames, adding/removing objects and changing scene geometry. For example, having seen a bus following a road, One-Shot GAN varies the length of a bus and placement of trees. In contrast, SinGAN, which is tuned to learn from a single image, does not generalize to the Single Video setting, producing "mean in class" textures and failing to learn appearance of objects (low diversity and very high SIFID). FastGAN, on the other hand, learns high-scale scene properties, but fails to augment the training set with non-trivial changes, having a very low distance to the training data (0.13 in Table 1).

Table 1 confirms that the proposed two-branch discriminator in combination with diversity regularization manages to overcome the memorization effect, achieving high distance to training data in both settings (0.37 and 0.34). This means that One-Shot GAN augments the training set with structural transformations that are orthogonal to standard data augmentation techniques, such as horizontal flipping or color jittering. To achieve this, the model requires as little data as one image or one short video clip. We believe, such ability can be especially useful to generate samples for augmentation of limited data, for example by creating new versions of rare examples.

4 Conclusion

We propose One-Shot GAN, a new unconditional generative model operating at different one-shot settings, such as learning from a single image or a single video. At such low-data regimes, our model mitigates the memorization problem and generates diverse images that are structurally different from the training set. We note that, inherently, one-shot image synthesis is constrained by the semantic diversity present in the original sample, e.g. for a video with a bus our model will not generate buses of different models. Nevertheless, our One-Shot GAN can synthesize images with novel views and different positions of the bus. We believe, such structural diversity provides a useful tool for augmentation in domains, where collecting data remains challenging.

REFERENCES

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems (NeurIPs)*, 2016.

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 (NeurIPS)*, 2014.

Tobias Hinz, Matthew Fisher, Oliver Wang, and Stefan Wermter. Improved techniques for training single-image gans. *arXiv preprint arXiv:2003.11512*, 2020.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Animesh Karnewar and Oliver Wang. Msg-gan: multi-scale gradient gan for stable image synthesis. *arXiv preprint arXiv:1903.06048*, 2019.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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)*, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020b.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. 2012.

Chuan Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision (ECCV)*, 2016.

Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: Generation by parts via conditional coordinating. In *International Conference on Computer Vision (ICCV)*, 2019.

Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized {gan} training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=1Fqg133qRaI`.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. *arXiv:1903.05854*, 2019.

E. Robb, Wensheng Chu, A. Kumar, and J. Huang. Few-shot adaptation of generative adversarial networks. *arXiv preprint arXiv:2010.11943*, 2020.

Edgar Schönfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Zhengsu Chen Jianwei Niu Qi Tian. Dropfilter: Dropout for convolutions. *arXiv preprint arXiv:1810.09849*, 2018.

Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 648–656, 2015.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, 2003.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Rui Xu, Xintao Wang, Kai Chen, Bolei Zhou, and Chen Change Loy. Positional encoding as spatial inductive bias in gans. *arXiv preprint arXiv:2012.05217*, 2020.

Dingdong Yang, Seunghoon Hong, Y. Jang, T. Zhao, and H. Lee. Diversity-sensitive conditional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019.

Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine learning (ICML)*, 2019.

Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans. *arXiv:2002.04724*, 2020a.

Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020b.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.