International Journal of Computer Vision (2022) 130:2903–2923
https://doi.org/10.1007/s11263-022-01673-x

OASIS: Only Adversarial Supervision for Semantic Image Synthesis

Vadim Sushko[1] · Edgar Schönfeld[1] · Dan Zhang[1,2] · Juergen Gall[3] · Bernt Schiele[4] · Anna Khoreva[1,2]

Received: 6 May 2022 / Accepted: 11 August 2022 / Published online: 17 September 2022

Abstract
Despite their recent successes, generative adversarial networks (GANs) for semantic image synthesis still suffer from poor image quality when trained with only adversarial supervision. Previously, additionally employing the VGG-based perceptual loss has helped to overcome this issue, significantly improving the synthesis quality, but at the same time limited the progress of GAN models for semantic image synthesis. In this work, we propose a novel, simplified GAN model, which needs only adversarial supervision to achieve high quality results. We re-design the discriminator as a semantic segmentation network, directly using the given semantic label maps as the ground truth for training. By providing stronger supervision to the discriminator as well as to the generator through spatially- and semantically-aware discriminator feedback, we are able to synthesize images of higher fidelity and with a better alignment to their input label maps, making the use of the perceptual loss superfluous. Furthermore, we enable high-quality multi-modal image synthesis through global and local sampling of a 3D noise tensor injected into the generator, which allows complete or partial image editing. We show that images synthesized by our model are more diverse and follow the color and texture distributions of real images more closely. We achieve a strong improvement in image synthesis quality over prior state-of-the-art models across the commonly used ADE20K, Cityscapes, and COCO-Stuff datasets using only adversarial supervision. In addition, we investigate semantic image synthesis under severe class imbalance and sparse annotations, which are common aspects in practical applications but were overlooked in prior works. To this end, we evaluate our model on LVIS, a dataset originally introduced for long-tailed object recognition. We thereby demonstrate high performance of our model in the sparse and unbalanced data regimes, achieved by means of the proposed 3D noise and the ability of our discriminator to balance class contributions directly in the loss function. Our code and pretrained models are available at https://github.com/boschresearch/OASIS.

Keywords Semantic image synthesis · GAN · Semantic segmentation · Label-to-image translation · Image editing

1 Introduction

Conditional generative adversarial networks (GANs) (Mirza & Osindero, 2014) synthesize images conditioned on class labels (Brock et al., 2019; Casanova et al., 2021), text (Reed et al., 2016; Zhang et al., 2018a, 2021), other images (Isola et al., 2017; Huang et al., 2018; Park et al., 2020), or semantic label maps (Park et al., 2019b; Liu et al., 2019; Wang et al., 2021b). In this work, we focus on the latter, addressing semantic image synthesis. Taking pixel-level annotated semantic maps as input, semantic image synthesis enables the rendering of realistic images from user-specified layouts,

Communicated by Arun Mallya.

Vadim Sushko and Edgar Schönfeld have contributed equally to this work.

✉ Vadim Sushko
vadim.sushko@bosch.com

Edgar Schönfeld
edgar.schoenfeld@bosch.com

Dan Zhang
dan.zhang2@bosch.com

Juergen Gall
gall@iai.uni-bonn.de

Bernt Schiele
schiele@mpi-inf.mpg.de

Anna Khoreva
anna.khoreva@bosch.com

[1] Bosch Center for Artificial Intelligence, Renningen, Germany

[2] University of Tübingen, Tübingen, Germany

[3] University of Bonn, Bonn, Germany

[4] Max Planck Institute for Informatics, Saarbrücken, Germany

Fig. 1 Existing semantic image synthesis models heavily rely on the VGG-based perceptual loss to improve the quality of generated images. In contrast, our model (OASIS) can synthesize diverse and high-quality images while only using an adversarial loss, without any external supervision

without the use of an intricate graphics engine. Therefore, its applications range widely from content creation and image editing to producing training data for downstream applications that adhere to specific semantic requirements (Park et al., 2019a; Ntavelis et al., 2020).

Despite the recent progress on stabilizing GANs (Miyato et al., 2018; Zhang & Khoreva, 2019; Karras et al., 2020a; Sauer et al., 2021) and developing their architectures (Karras et al., 2021, 2019, 2020b; Brock et al., 2019; Liu et al., 2021), state-of-the-art GAN-based semantic image synthesis models (Park et al., 2019b; Liu et al., 2019; Wang et al., 2021b) still greatly suffer from training instabilities and poor image quality when the generator is only trained to fool the discriminator in an adversarial fashion (see Fig. 1). An established practice to overcome this issue is to employ a perceptual loss (Wang et al., 2018) to train the generator, in addition to the discriminator loss. The perceptual loss aims to match intermediate features of synthetic and real images, that are estimated via an external perception network. A popular choice for such a network is VGG (Simonyan & Zisserman, 2015), pre-trained on ImageNet (Deng et al., 2009). Although the perceptual loss substantially improves the performance of previous methods, it comes with the computational overhead introduced by utilizing an extra network for training. Moreover, as we show in our experiments, it dominates over the adversarial loss during training, as the generator starts to learn mostly through minimizing the VGG loss, which has a negative impact on the diversity and quality of generated images. Therefore, in this work we propose a novel, simplified model that establishes new state-of-the-art results without requiring a perceptual loss.

To achieve semantic image synthesis of high quality, the training signal to the GAN generator should contain feedback on whether the generated images are well aligned to the input label maps. Thus, a fundamental question for GAN-based semantic image synthesis models is how to design the discriminator that would efficiently utilize information from given semantic label maps, in addition to judging the realism of given images. Conventional methods (Park et al., 2019b; Wang et al., 2018, 2021b; Liu et al., 2019; Isola et al., 2017; Ntavelis et al., 2020) adopt a multi-scale classification network, taking the label map as input along with the image, and making a global image-level real/fake decision. This discriminator has limited representation power, as it is not incentivized to learn high-fidelity pixel-level details of the images and their precise alignment with the input semantic label maps. For example, such a classification-based discriminator can base its decision solely on image realism, without the need of examining the alignment between the image and label map. To mitigate this issue, we propose an alternative architecture for the discriminator, re-designing it as an encoder-decoder semantic segmentation network (Ronneberger et al., 2015), and directly exploiting the given semantic label maps as ground truth via an $(N + 1)$-class cross-entropy loss. This new discriminator provides semantically-aware pixel-level feedback to the generator, partitioning the image into segments belonging to one of the N real semantic classes or the fake class. With this design, the network cannot ignore the provided label maps, as it has to predict a correct class label for each pixel of an image. Enabled by the discriminator per-pixel response, we further introduce a LabelMix regularization, which fosters

Fig. 2 OASIS multi-modal synthesis results. The 3D noise can be sampled globally (first 2 rows), changing the whole scene, or locally (last 2 rows), partially changing the image. For the latter, we sample different noise per region, like the bed segment (in red) or arbitrary areas defined by shapes (Color figure online)

the discriminator to focus more on the semantic and structural differences of real and synthetic images. The proposed changes lead to a much stronger discriminator, that maintains a powerful semantic representation of objects, giving more meaningful feedback to the generator, and thus making the perceptual loss supervision superfluous (see Fig. 1).

Semantic image synthesis is naturally a one-to-many mapping, where one label map can correspond to many possible real images. Thus, a desirable property of a generator is to generate a diverse set of images from a single label map, only by sampling noise. This property is known as multi-modality. Previously, only using a noise vector as input was not sufficient to achieve multi-modality, because the generator tended to mostly ignore the noise or synthesized images of poor quality (Isola et al., 2017; Wang et al., 2018). Thus, prior work (Wang et al., 2018; Park et al., 2019b) resorted to using an image encoder to produce multi-modal outputs. In this work, we enable multi-modal synthesis of the generator via a newly-introduced 3D noise sampling method, without requiring an image encoder and not relying on availability of a reference image to produce new image styles. Empowered by our stronger discriminator, the generator can now effectively synthesize different images by simply resampling a 3D noise tensor, which is used not only as the input, but is also combined with intermediate features via conditional normalization at every layer. This procedure makes the generator spatially sensitive to noise, so we can re-sample it both globally (channel-wise) and locally (pixel-wise), allowing to change not only the appearance of the whole scene, but also of specific semantic classes or any chosen area (see Fig. 2). As shown in our experiments, the proposed 3D noise injection scheme enables a significantly higher diversity of synthesis compared to previous methods.

With the proposed modifications in the discriminator and generator design, we outperform the prior state of the art in synthesis quality across the commonly used ADE20K (Zhou et al., 2017), COCO-Stuff (Caesar et al., 2018) and Cityscapes (Cordts et al., 2016) datasets. Omitting the necessity of the VGG perceptual loss, our model generates samples of higher quality and diversity, and follows the color and texture distributions of real images more closely.

A well known challenge for semantic segmentation applications is the problem of class imbalance. In practice, a dataset can contain underrepresented classes (representing a very small fraction of the dataset pixels), which can lead to suboptimal performance of models (Sudre et al., 2017). However, to the best of our knowledge, this problem has not been studied in the context of semantic image synthesis. For this reason, we propose to extend the evaluation setup used in previous works by using the highly imbalanced LVIS dataset (Gupta et al., 2019). Originally introduced as a dataset for long-tailed object recognition, LVIS contains

a large set of 1203 classes, the majority of which appear only in a few images. Moreover, to simplify dataset curation, label maps in LVIS were annotated sparsely, with large image areas being occupied with a generic background label. The above properties make LVIS a very challenging evaluation setting for previous semantic image synthesis models, as we demonstrate by the example of the state-of-the-art SPADE model (Park et al., 2019b). As the classification-based discriminator of SPADE makes a global real/fake decision for each image-label pair, the loss contribution originating from underrepresented classes can be dominated by the loss contribution of well represented classes. In contrast, our proposed discriminator mitigates this issue: with the $(N + 1)$-class cross-entropy loss computed for each image pixel, it becomes possible to assign higher weights for the pixels belonging to underrepresented classes. As shown in our experiments, our model successfully deals with both the extreme class imbalance and sparsity in label maps, outperforming SPADE on the LVIS dataset by a large margin.

To extend the evaluation of our model further, we test the efficacy of generated images when applied as synthetic data augmentation for the training of semantic segmentation networks. This way, the performance of semantic image synthesis is assessed through a task that holistically requires high image quality, diversity, and precise image alignment to the label maps. We demonstrate that the synthetic data produced by our model achieves high performance on this test, eliciting a notable increase in downstream segmentation performance. In doing so, our model outperforms a strong baseline SPADE (Park et al., 2019b), indicating its high potential to be applied in segmentation applications. In addition, we also demonstrate how our model for the first time enables the application of a GAN-based semantic image synthesis model to unlabelled images, without requiring external segmentation networks. Thanks to a good segmentation performance of our trained discriminator, we can infer the label map of an image and generate many alternative versions of the same scene by varying the 3D noise. We find these results promising for future utilization of our model in applications.

We call our model OASIS, as it needs **o**nly **a**dversarial supervision for **s**emantic **i**mage **s**ynthesis. In summary, our main contributions include:

- We propose a novel segmentation-based discriminator architecture, that gives more powerful feedback to the generator and eliminates the necessity of the perceptual loss supervision.
- We present a simple 3D noise sampling scheme, notably increasing the diversity of multi-modal synthesis and enabling both complete or partial resampling of a generated image.
- With the OASIS model, we achieve high-quality results on the ADE20K, Cityscapes and COCO-Stuff datasets, outperforming previous state-of-the-art models while relying only on adversarial supervision. We show that images synthesized by OASIS exhibit much higher diversity and more closely follow the color and texture distributions of real images.
- We propose to use the LVIS dataset (Gupta et al., 2019) to assess image generation in the regime with many underrepresented semantic classes, leading to a severe class imbalance. We show how the OASIS design directly addresses these issues and thereby outperforms the strong baseline SPADE (Park et al., 2019b) by a large margin.
- We test the efficacy of generated images for synthetic data augmentation, as a unified measure that simultaneously depends on image quality, diversity, and label map alignment. The images generated by OASIS elicit a stronger increase in downstream segmentation performance compared to SPADE, suggesting a higher potential of our model for future utilization in applications.

This paper is an extended version of our previous work (Schönfeld et al., 2021). Compared to the prior conference version, we provide a a significantly extended experimental evaluation and a more in-depth discussion of the proposed contributions. In particular, the conventional evaluation setup is extended to the extremely imbalanced data regime on the LVIS dataset (see Sect. 4.3). We further extend the evaluation by testing the efficacy of synthetic images as data augmentation for the task of semantic segmentation (see Sect. 4.5). We add new results on the synthesis of diverse images from unlabelled data (see Sect. 4.4 and Fig. 13). These new results highlight specific benefits of our approach compared to other models. Finally, we offer a new detailed ablation study of the method (see Tables 7, 10, 11, 12a) and extend the discussion of our model by analysing its independence on the perceptual loss (Sect. 3.4).

2 Related Work

Semantic image synthesis. The task of semantic image synthesis is to solve the inverse problem of semantic image segmentation: generate photorealistic and diverse images from provided semantic label maps. Currently, the most prominent approaches for this task are based on conditional GANs (Mirza & Osindero, 2014), as first proposed by the Pix2pix model (Isola et al., 2017). Pix2pix generates images with an encoder-decoder generator that takes label maps as input, and employs a PatchGAN discriminator which is induced to distinguish between real and fake image-label pairs. Lately, various GAN models with modified generator and discriminator architectures have been introduced (Wang et al., 2018; Park et al., 2019b; Liu et al., 2019; Tang et al., 2020c, b; Ntavelis et al., 2020; Wang et al., 2021b; Richard-

son et al., 2021; Li et al., 2021) to improve the quality and diversity of image synthesis. Besides GANs, Chen and Koltun (2017) proposed to use a cascaded refinement network (CRN) for high-resolution semantic image synthesis, and SIMS (Qi et al., 2018) extended it with a non-parametric component, serving as a memory bank of source material to assist the synthesis. Further, Li et al. (2019) employed implicit maximum likelihood estimation (Li & Malik, 2018) to increase the synthesis diversity of the CRN model. However, these approaches still underperform in comparison to state-of-the-art GAN models. Therefore, we next focus on the recent GAN architectures for semantic image synthesis.

Discriminator architectures. To provide a powerful guiding signal to the generator, a GAN discriminator for semantic image synthesis should evaluate both the image realism and its alignment to the provided semantic label map. Thus, a fundamental question is to find the most efficient way for the discriminator to utilize the given semantic label maps. To this end, Pix2pix (Isola et al., 2017), Pix2pixHD (Wang et al., 2018) and SPADE (Park et al., 2019b) rely on concatenating the label maps directly to the input image, which is fed to a multi-scale PatchGAN discriminator. Alternatively, SESAME (Ntavelis et al., 2020) employed a projection-based discriminator (Miyato & Koyama, 2018), applying an additional branch to process semantic label maps separately from images, and merging the two streams before the last convolutional layer via a pixel-wise multiplication. CC-FPSE (Liu et al., 2019) proposed a feature-pyramid discriminator, embedding both images and label maps into a joint feature map, and then consecutively upsampling it in order to classify it as real/fake at multiple scales. LGGAN (Tang et al., 2020c) introduced a classification-based feature learning module to learn more discriminative and class-specific features. In this work, we propose to use a simple pixel-wise semantic segmentation network as a discriminator instead of multi-scale image classifiers as in the above approaches, and to directly exploit the semantic label maps for its supervision. Segmentation-based discriminators have been shown to improve semantic segmentation (Souly et al., 2017) and unconditional image synthesis (Schönfeld et al., 2020), but to the best of our knowledge have not been explored for semantic image synthesis and our work is the first to apply an adversarial semantic segmentation loss for this task.

Generator architectures. To enforce the alignment between the generated images and the conditioning label maps, previous methods explored different ways to incorporate the label maps into the generator training. In many conventional approaches (Isola et al., 2017; Wang et al., 2018; Tang et al., 2020b, c; Ntavelis et al., 2020; Richardson et al., 2021), label maps are provided to the generator via an additional encoder network. However, this solution has been shown to be suboptimal at preserving the semantic information until the later stages of image generation. Therefore, SPADE introduced a spatially-adaptive normalization layer that directly modulates the label map onto the generator's hidden layer outputs at various scales. Alternatively, CC-FPSE proposed to use spatially-varying convolution kernels conditioned on the label map. Most recently, SC-GAN (Wang et al., 2021b) utilized label maps as input to generate class-specific semantic vectors at different scales, which are used as conditioning at different layers of the image rendering network; and CollageGAN (Li et al., 2021) proposed to extract a label map representation via feature pyramid encoder and inject it as spatial style tensor to a StyleGAN2 generator.

While improving the quality of generated images, the above models struggled to achieve multi-modality through sampling the input noise, as the generator tended to become insensitive to noise or achieved only poor quality, as first observed by (Isola et al., 2017). Thus, the above approaches resorted to having an image encoder in the generator design to enable multi-modal synthesis. The generator then combines the extracted image style with the label map to reconstruct the original image. By alternating the style vector, one can generate multiple outputs conditioned on the same label map. However, using an image encoder is a resource-demanding solution. In this work, we enable multi-modal synthesis directly through sampling of a 3D noise tensor which is injected at every layer of the network. Different from the structured noise injection of Alharbi and Wonka (2020) and class-specific latent codes of Zhu et al. (2020), we inject the 3D noise along with label maps and adjust it to image resolution, also enabling re-sampling of selected semantic segments (see Fig. 2).

Perceptual losses. Gatys et al. (2015, 2016); Johnson et al. (2016) and Bruna et al. (2016) were pioneers at exploiting perceptual losses to produce high-quality images for super-resolution and style transfer using convolutional networks. Such a loss extracts deep features from real and generated images by an external classification network, and minimizes their L1-distance to bring fake images closer to the real data. For semantic image synthesis, the VGG-based perceptual loss was first introduced by CRN (Chen & Koltun, 2017), and later adopted by Pix2pixHD (Isola et al., 2017). Since then, it has become a default for training the generator (Park et al., 2019b; Liu et al., 2019; Tan et al., 2020; Tang et al., 2020a; Richardson et al., 2021; Wang et al., 2021b; Li et al., 2021). As the perceptual loss is based on a VGG network pre-trained on ImageNet (Deng et al., 2009), methods relying on it are constrained by the ImageNet domain and the representational power of VGG. With the recent progress on GAN training, e.g., by architecture designs and regularization techniques, the actual necessity of the perceptual loss requires a reassessment. We experimentally show that such loss imposes unnecessary constraints on the generator, significantly limiting the diversity among samples. Trained without

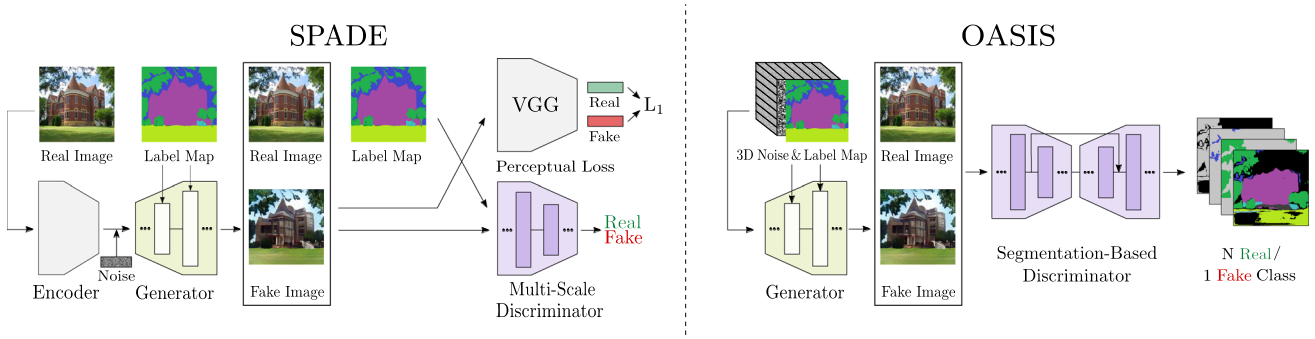

Fig. 3 SPADE (left) versus OASIS (right). OASIS outperforms SPADE, while being simpler and lighter: it uses only an adversarial loss as supervision and a single segmentation-based discriminator, without relying on heavy external networks. Furthermore, OASIS learns to synthesize multi-modal outputs by directly re-sampling the 3D noise tensor, instead of using an image encoder as in SPADE

the VGG loss, our model achieves improved diversity, at the same time not compromising the quality of generated images.

3 The OASIS Model

In this section, we present our OASIS model, which, in contrast to other semantic image synthesis methods, needs only adversarial supervision for training. Using SPADE as a starting point (Sect. 3.1), we first propose to re-design the discriminator as a semantic segmentation network, directly using the given semantic label maps as ground truth (Sect. 3.2). Empowered by spatially- and semantically-aware feedback of the new discriminator, we next re-design the SPADE generator, enabling its effective multi-modal synthesis via 3D noise sampling (Sect. 3.3). Lastly, we illustrate the superfluity of the VGG loss for our model (Sect. 3.4).

3.1 The SPADE Baseline

We choose SPADE as our baseline as it is a state-of-the-art model and a relatively simple representative of conventional semantic image synthesis models. As depicted in Fig. 3, the discriminator of SPADE largely follows the PatchGAN multi-scale discriminator (Isola et al., 2017), adopting two image classification networks operating at different resolutions. Both of them take the channel-wise concatenation of the semantic label map and the real/fake image as input, and produce real/fake classification scores. On the generator side, SPADE adopts spatially-adaptive normalization layers to effectively integrate the semantic label map into the synthesis process from low to high scales. Additionally, the image encoder is used to extract the style vector from the reference image, which is then combined with a 1D noise vector for multi-modal synthesis. The training loss of SPADE consists of three terms, namely, an adversarial loss, a feature matching loss and the VGG-based perceptual loss:

$$\mathcal{L} = \max_G \min_D \mathcal{L}_{\text{adv}} + \lambda_{\text{fm}} \mathcal{L}_{\text{fm}} + \lambda_{\text{vgg}} \mathcal{L}_{\text{vgg}}. \quad (1)$$

Overall, SPADE is a resource-demanding model at both training and test time, i.e., with two PatchGAN discriminators, an image encoder in addition to the generator, and the VGG loss. In the following, we revisit its architecture and introduce a simpler and more efficient solution that offers better performance and reduces the model complexity.

3.2 The OASIS Discriminator

To train the generator to synthesize high-quality images that are well aligned with the input semantic label maps, we need a powerful discriminator that coherently captures discriminative semantic features at different image scales. While classification-based discriminators, such as PatchGAN, take label maps as input concatenated to images, they can afford to ignore them and make the decision solely on image patch realism. Thus, we propose to cast the discriminator task as a multi-class semantic segmentation problem to directly utilize label maps for supervision, and accordingly alter its architecture to an encoder-decoder segmentation network (see Fig. 3). Encoder-decoder networks have proven to be effective for semantic segmentation (Badrinarayanan et al., 2016; Chen et al., 2018). Thus, we build our discriminator architecture upon U-Net (Ronneberger et al., 2015), which consists of the encoder and decoder connected by skip connections. This discriminator architecture is multi-scale through its design, integrating information over up- and down-sampling pathways as well as through the encoder-decoder skip connections. The segmentation task of the discriminator is formulated to predict the per-pixel class label of the real images, using the given semantic label maps as ground truth. In addition to the N semantic classes from the label maps, all pixels of fake images are categorized as one extra class. As the formulated semantic segmentation problem has $N + 1$ classes, we propose to use an $(N + 1)$-class cross-entropy loss for training.

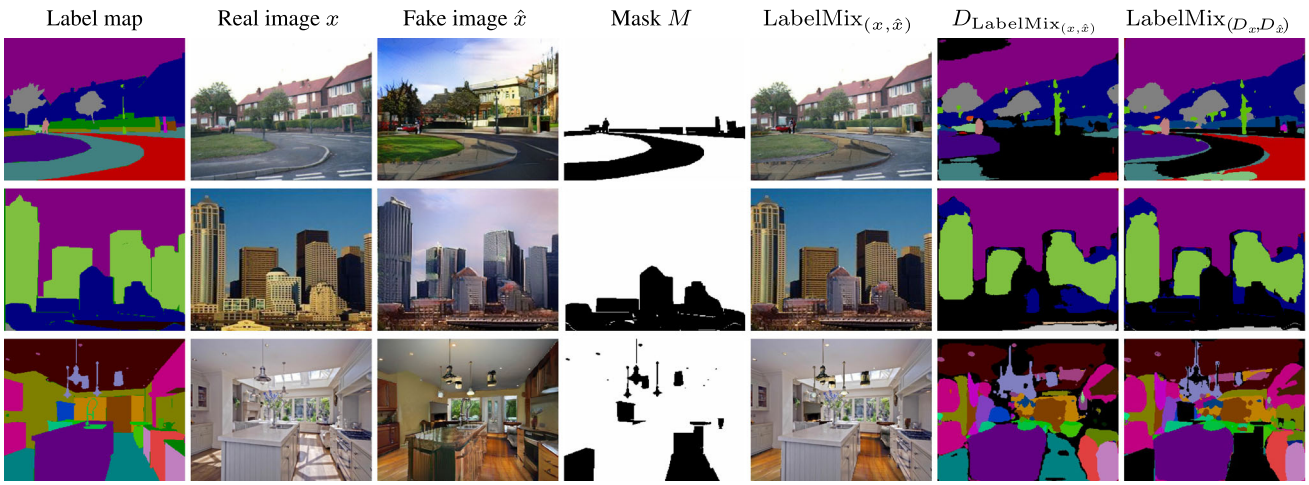

Fig. 4 LabelMix regularization. Real x and fake $\hat{x}$ images are mixed using a binary mask M, sampled based on the label map, resulting in LabelMix$_{(x,\hat{x})}$. The consistency regularization minimizes the L2 distance between the logits of $D_{\text{LabelMix}_{(x,\hat{x})}}$ and LabelMix$_{(D_x,D_{\hat{x}})}$. In this visualization, **black** corresponds to the fake class in the $N+1$ segmentation output

In practice, the N semantic classes are often imbalanced, as some of the classes represent significantly less pixels of the dataset compared to others. The loss contribution for such underrepresented classes can be dominated by well represented classes, which can lead to suboptimal performance. To mitigate this issue, empowered by the pixel-level loss computation of our discriminator, we propose to weight each class by its inverse pixel-wise frequency in a batch, thus giving underrepresented semantic classes more weight. In doing so, the loss contributions of each class are equally balanced, and, thus, the generator is also encouraged to pay more attention to underrepresented classes. Mathematically, the new discriminator loss is expressed as:

$$\begin{aligned}\mathcal{L}_D = &- \mathbb{E}_{(x,t)}\left[\sum_{c=1}^{N}\alpha_c \sum_{i,j}^{H\times W} t_{i,j,c}\log D(x)_{i,j,c}\right] \\ &- \mathbb{E}_{(z,t)}\left[\sum_{i,j}^{H\times W}\log D(G(z,t))_{i,j,c=N+1}\right],\end{aligned} \tag{2}$$

where x denotes the real image; (z, t) is the noise-label map pair used by the generator G to synthesize a fake image; and the discriminator D maps the real or fake image into a per-pixel $(N + 1)$-class prediction probability. The ground truth label map t has three dimensions, where the first two correspond to the spatial position $(i, j) \in H \times W$, and the third one is a one-hot vector encoding the class $c \in \{1, .., N+1\}$. The class balancing weight α_c is the inverse pixel-wise frequency of a class c per batch:

$$\alpha_c = \frac{H \times W}{\sum_{i,j}^{H\times W} E_t\left[\mathbb{1}[t_{i,j,c}=1]\right]}. \tag{3}$$

In effect, improving the synthesis of underrepresented and well represented classes is equally necessary to minimize the loss. As we show in Sect. 4.3, this step helps to improve the synthesis quality of underrepresented classes.

LabelMix regularization. In order to encourage our discriminator to focus on differences in content and structure between the fake and real classes, we propose a LabelMix regularization. Based on the semantic layout, we generate a binary mask M to mix a pair $(x, \hat{x})$ of real and fake images conditioned on the same label map: $\text{LabelMix}(x, \hat{x}, M) = M \odot x + (1 - M) \odot \hat{x}$, as visualized in Fig. 4. Given the mixed image, we further train the discriminator to be equivariant under the LabelMix operation. This is achieved by adding a consistency loss term $\mathcal{L}_{cons}$ to Eq. 2:

$$\begin{aligned}\mathcal{L}_{cons} = &\Big\| D_{\text{logits}}\Big(\text{LabelMix}(x, \hat{x}, M)\Big) \\ &- \text{LabelMix}\Big(D_{\text{logits}}(x), D_{\text{logits}}(\hat{x}), M\Big)\Big\|^2,\end{aligned} \tag{4}$$

where D_{logits} are the logits attained before the last softmax activation layer, and $\|\cdot\|$ is the L_2 norm. This consistency loss compares the output of the discriminator on the LabelMix image with the LabelMix of its outputs, penalizing the discriminator for inconsistent predictions. LabelMix is different to CutMix (Yun et al., 2019), which randomly samples the binary mask M. A random mask will introduce inconsistency between the pixel-level labels and the scene layout provided by the label map. For an object with the class label c, it will contain pixels from both real and fake images, resulting in two labels, i.e. c and $N+1$. To avoid such inconsistency, the mask of LabelMix is generated according to the label map, providing natural borders between semantic regions, see Mask M in Fig. 4. Under LabelMix regular-

ization, the generator is encouraged to respect the natural semantic boundaries, improving pixel-level realism while also considering the class segment shapes.

Alternative ways to encode label maps. Besides the proposed $(N+1)$-class cross entropy loss, there are other ways to incorporate a label map into the training of a segmentation-based discriminator. One can concatenate the label map to the input image, analogous to SPADE. Another option is to use projection, by taking the inner product between the last linear layer output and the embedded label map, analogous to class-label conditional GANs (Miyato & Koyama, 2018). For both alternatives, the training loss is the pixel-level real/fake binary cross-entropy (Schönfeld et al., 2020). As in these two variants the label maps are used as input to the discriminator (concatenated to the input image or fed to the last linear layer), they are propagated *forward* through the network. In contrast, the (N+1)-setting uses label maps only as targets for the loss computation, so they are propagated *backward* through the network via the gradients updates. Backward propagation ensures that the discriminator learns semantic-aware features, in contrast to forward propagation, where the alignment of a generated image to the input label map can be ignored. The comparison between the above label map encodings is shown in Table 9.

3.3 The OASIS Generator

To stay in line with the OASIS discriminator design, the training loss for the generator is changed to

$$\mathcal{L}_G = -\mathbb{E}_{(z,t)}\left[\sum_{c=1}^{N}\alpha_c \sum_{i,j}^{H\times W} t_{i,j,c}\log D(G(z,t))_{i,j,c}\right], \quad (5)$$

which is a direct outcome of the non-saturation trick (Goodfellow et al., 2014) to Eq. 2. We next re-design the generator to enable multi-modal synthesis through noise sampling. SPADE is deterministic in its default setup, but can be trained with an extra image encoder to generate multi-modal outputs. We introduce a simpler version, that enables synthesis of diverse outputs directly from input noise. For this, we construct a noise tensor of size $M\times H\times W$, matching the spatial dimensions of the label map of size $N\times H\times W$, where N is the number of semantic labels and $H\times W$ corresponds to the height and width of the image. Note that for simplicity during training we sample the 3D noise tensor globally, i.e. per-channel, replicating each channel value spatially along the height and width of the tensor. In other words, a M-dimensional latent vector is sampled and then broadcasted to each pixel of an image. We analyze alternative ways of sampling 3D noise during training in the ablation section (see Sect. 4.6). After sampling, the noise and the label map are concatenated along the channel dimensions to form a combined noise-label 3D tensor of size $(M+N)\times H\times W$. This combined tensor serves as input to the first generator layer, but also as input to the spatially-adaptive normalization layers in every generator block. This way, all intermediate feature maps are conditioned on both the semantic labels and the noise (see Fig. 3), making the noise hard to ignore. As the 3D noise is channel- and pixel-wise sensitive, at test time, one can sample the noise globally, per-channel, and locally, per-segment or per-pixel, for controlled synthesis of the whole scene or of specific semantic objects. For example, when generating a scene of a bedroom, one can re-sample the noise locally and change the appearance of the bed alone (see Fig. 2).

Note that using image styles via an encoder, as in SPADE, is also possible in our setting, as the 3D noise can be simply concatenated to the encoder style features. Lastly, to further reduce the complexity, we remove the first residual block in the generator, reducing the number of parameters from 96M to 72M without a noticeable performance loss (see Table 7).

3.4 Superfluity of the Perceptual Loss for OASIS

In contrast to SPADE, which strongly relies on the perceptual loss during training (see Fig. 1), the OASIS generator is trained only with the adversarial loss from the segmentation-based discriminator, according to Eq. 5. To illustrate the insignificance of the VGG loss for OASIS, in Fig. 5 we compare the curves of the VGG and generator adversarial loss functions of SPADE and OASIS, for comparison additionally trained with the perceptual loss. We see that SPADE focuses on minimizing the VGG loss during training, but keeps the adversarial generator loss constant. Without a rich training signal from its Patch-GAN discriminator, the generator of SPADE resorts to learning mostly from the VGG loss. In contrast, with the stronger discriminator supervision provided by the semantic label maps and the multi-scale U-Net architecture, OASIS achieves a better adversarial balance. Hence, the generator is forced to learn semantically meaningful features that the segmentation-based discriminator judges as real, and the generator loss does not stay constant (see Fig. 5).

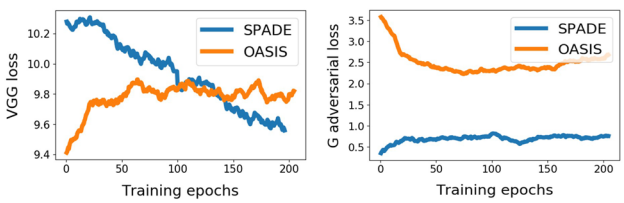

Fig. 5 VGG and adversarial generator loss functions for SPADE and OASIS trained with VGG loss on ADE20k dataset. The adversarial loss scales are different due to different objectives (binary or (N+1)-class cross entropy loss)

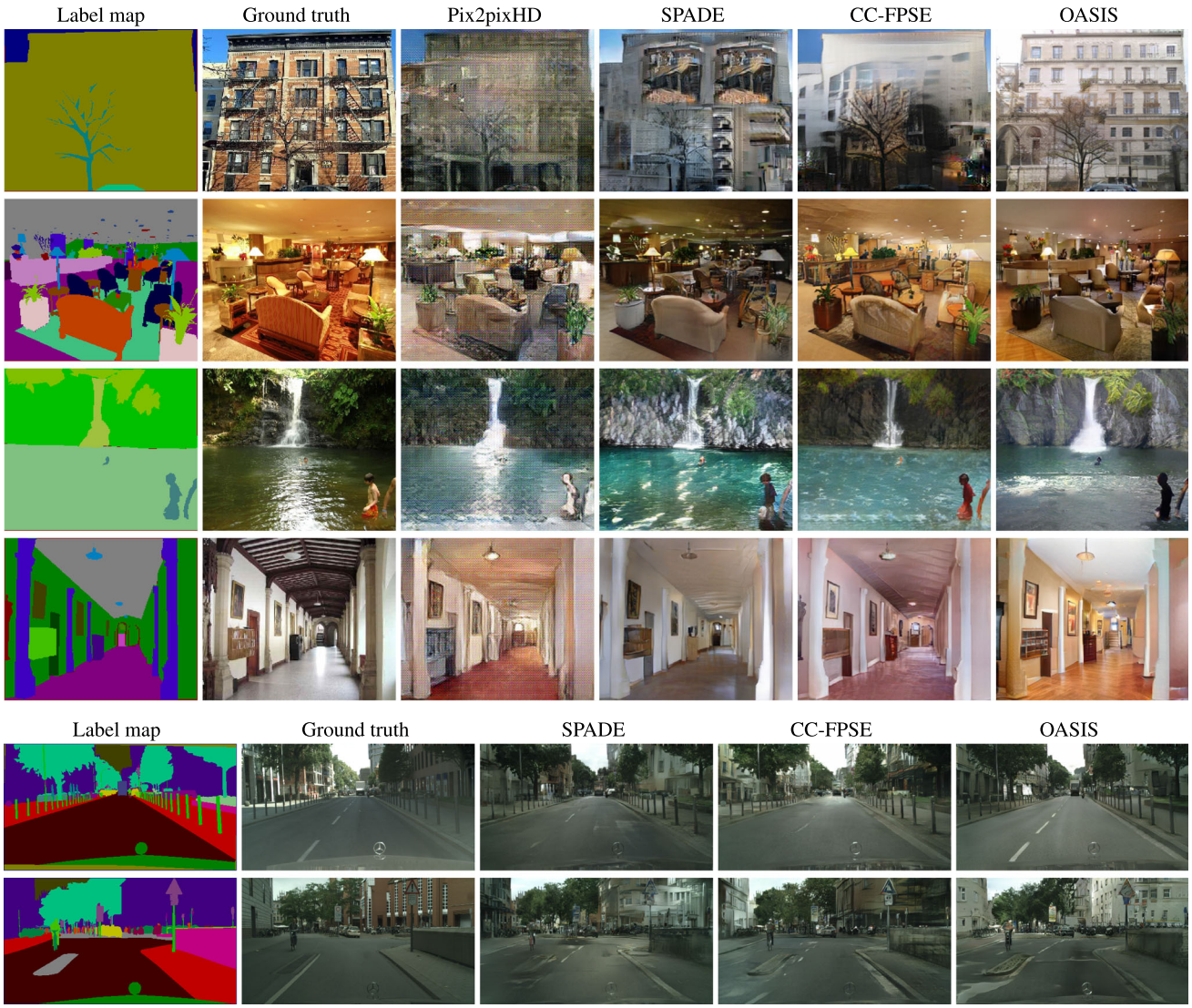

Fig. 6 Qualitative comparison of OASIS with other methods on ADE20K and Cityscapes. Trained with only adversarial supervision, our model generates images with better perceptual quality and structure

The advantage of training the generator only with the adversarial loss is three-fold. Firstly, the perceptual loss can bias the training signal with the color and texture statistics encoded in the VGG features extracted from ImageNet. As shown in Sect. 4.2, the strong adversarial supervision from the OASIS discriminator, without the VGG loss, allows to generate images with color and texture distributions closer to the provided real data. Secondly, the perceptual loss can induce unnecessary constraints on the generator and thus significantly limit the diversity of multi-modal image syntesis. This effect is further demonstrated in Table 2. Lastly, removing the perceptual loss eliminates the computational overhead which was introduced by an additional VGG network during training.

4 Experiments

We provide an extensive experimental evaluation of our contributions, using the official implementation of SPADE[1] as our baseline. The setup of our experiments is described in detail in Sect. 4.1. Firstly, we compare OASIS with prior methods on common semantic image synthesis benchmark datasets, comparing their performance in terms of both image quality and diversity (Sect. 4.2). To further highlight the advantages of OASIS over the SPADE baseline, we provide additional discussions on different aspects of the semantic image synthesis. In particular, Sect. 4.3 is devoted to the performance analysis on the underrepresented classes, extending the comparison of the models to the LVIS

[1] https://github.com/NVlabs/SPADE.

dataset (Gupta et al., 2019). Section 4.4 demonstrates new semantic image editing techniques enabled by OASIS. Section 4.5 explores the application of generated images as synthetic data augmentation for the training of semantic segmentation networks. Lastly, we provide an extensive ablation study to verify the effectiveness of the proposed contributions (Sect. 4.6).

4.1 Experimental Setup

Datasets. We conduct experiments on several challenging datasets. Firstly, to compare OASIS with prior models, we use the ADE20K (Zhou et al., 2017), COCO-Stuff (Caesar et al., 2018) and Cityscapes (Cordts et al., 2016), which are the three benchmark datasets commonly used in the semantic image synthesis literature (see Sect. 4.2). The image resolution is set to 256x256 for ADE20K and COCO-Stuff, and 256x512 for experiments on Cityscapes. Following Qi et al. (2018), we also evaluate OASIS on ADE20K-outdoors, the subset of ADE20K containing only outdoor scenes.

Secondly, to test the capability of models to learn underrepresented classes, we conduct additional evaluations on the ADE20K and LVIS dataset (Gupta et al., 2019) (see Sect. 4.3). We select ADE20K among conventional datasets for its notable class imbalance, as among its 150 classes, more than 86% of the image pixels belong only to the 30 best represented ones (see Table 3). In addition, to test the networks under more extreme class imbalance, we propose to use LVIS, the dataset that has been originally introduced for the task of long-tailed instance segmentation. LVIS employs the same set of training images as COCO-Stuff, but its annotations are different in two important ways. Firstly, LVIS provides a significantly larger set of 1203 annotated classes, following a long-tailed distribution in which some classes are present only in one or a few training samples (see Fig. 7). Secondly, due to a fixed labelling budget, different background types were not considered for annotation in LVIS. Consequently, the images in LVIS dataset contain large areas belonging to the background class, which sometimes covers more than 90% of the pixels in an image (see grey areas in Fig. 10). For the above two reasons, the structure of LVIS poses a new challenge for semantic image synthesis, as models need to account for a much more extreme class imbalance. We conduct experiments on LVIS at the image resolution of 128x128.

Training. We follow the experimental setting of Park et al. (2019b). The Adam (Kingma & Ba, 2015) optimizer was used with momenta $\beta = (0, 0.999)$ and constant learning rates $(0.0001, 0.0004)$ for G and D. We did not use the GAN feature matching loss for OASIS, as we did not observe any improvement with it, and used the VGG loss only for ablations with $\lambda_{\text{VGG}} = 10$. The parameter for LabelMix λ_{LM} was set to 5 for ADE20k and Cityscapes, and to 10 for COCO-Stuff and LVIS. The latent dimension M was set to 64. We did not experience any training instabilities and, thus, did not employ any extra stabilization techniques. All our models use an exponential moving average (EMA) of the generator weights with 0.9999 decay. All the experiments were run on 4 Tesla V100 GPUs, with a batch size of 20 for Cityscapes and 32 for the other datasets. The training epochs are 200 on ADE20K and Cityscapes, and 100 for the larger COCO-Stuff and LVIS datasets. On average, a complete forward-backward pass with batch size 32 on ADE20k takes around 0.95ms per training image.

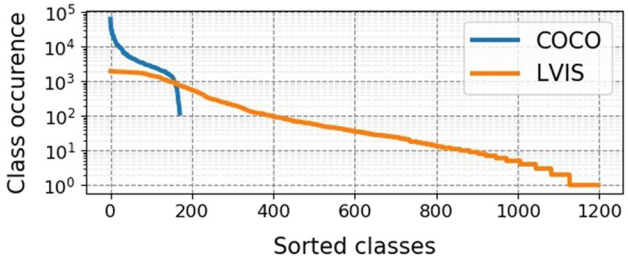

Fig. 7 Comparison of class distributions of the COCO and LVIS datasets. LVIS has a much larger vocabulary of 1203 classes with a long tail of underrepresented classes

Evaluation metrics. Following prior work (Park et al., 2019b; Liu et al., 2019), we evaluate the *quality* of semantic image synthesis by computing the FID (Heusel et al., 2017) and evaluate the *alignment* of the generated images with their semantic label maps via mIoU (mean intersection-over-union) or mAP (mean average precision) on the test set (see Sect. 4.2). mIoU evaluates the alignment of generated images with their ground truth label maps, as measured by an external pre-trained semantic segmentation network. We use UperNet101 (Xiao et al., 2018) for ADE20K, multi-scale DRN-D-105 (Yu et al., 2017) for Cityscapes, and DeepLabV2 (Chen et al., 2015) for COCO-Stuff. Differently, for the LVIS dataset, the alignment of generated images to ground truth label maps is measured using mAP instead of mIoU, following the official guidelines for evaluating instance segmentation models on this dataset (see Sect. 4.3). We compute mAP using a state-of-the-art instance segmentation model from Wang et al. (2021a), pre-trained on LVIS.

In addition, to better understand how the perceptual loss influences the synthesis performance, we propose to compare the *color and texture statistics* of generated and real images. For this, we compute color histograms in the LAB space and measure the earth mover's distance between the real and generated image sets (Rubner et al., 2000). We also measure the texture similarity to the real data as the χ^2-distance between Local Binary Patterns histograms (Ojala et al., 1996). As different semantic classes have different color and texture distributions, we aggregate the histogram distances separately per class and compute their average.

Table 1 Comparison with other methods across datasets

Method	# param	VGG	ADE20K		ADE-outd.		Cityscapes		COCO-stuff	
			FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑
CRN	84M	✓	73.3	22.4	99.0	16.5	104.7	52.4	70.4	23.7
SIMS	56M	✓	n/a	n/a	67.7	13.1	49.7	47.2	n/a	n/a
Pix2pixHD	183M	✓	81.8	20.3	97.8	17.4	95.0	58.3	111.5	14.6
LGGAN	n/a	✓	31.6	41.6	n/a	n/a	57.7	68.4	n/a	n/a
CC-FPSE	131M	✓	31.7	43.7	n/a	n/a	54.3	65.5	19.2	41.6
SC-GAN	66M	✓	29.3	45.2	n/a	n/a	49.5	66.9	18.1	42.0
SESAME	104M	✓	31.9	**49.0**	n/a	n/a	54.2	66.0	n/a	n/a
SPADE	102M	✓	33.9	38.5	63.3	30.8	71.8	62.3	22.6	37.4
SPADE+	102M	✓	32.9	42.5	51.1	32.1	47.8	64.0	21.7	38.8
		✗	60.7	21.0	65.4	22.7	61.4	47.6	99.1	16.1
OASIS	94M	✗	**28.3**	48.8	**48.6**	**40.4**	**47.7**	**69.3**	**17.0**	**44.1**

Bold denotes the best performance

To measure the *diversity* among synthesized samples in the multi-modal image generation regime, we evaluate MS-SSIM (Wang et al., 2003) and LPIPS (Zhang et al., 2018b) between the images generated from the same label map. For each label map in the test set, we generate 20 images and compute the mean pairwise scores. For the final numbers, the scores are averaged over all label maps.

Lastly, we propose to test the efficacy of generated images when applied as *synthetic data augmentation* for the task of semantic segmentation (see Sect. 4.5). For this, we take a DeepLab-V3 segmentation network with a ResNeSt-50 backbone (Zhang et al., 2020) and train it on ADE20K and Cityscapes. At each training step of DeepLab-V3, we add for each training image its synthetic counterpart to the batch, generated from the same label map. The efficacy of synthetic images is therefore measured by its effect on the downstream mIoU performance of DeepLab-V3.

4.2 Evaluation of the Synthesis Quality and Diversity

In this section, we compare OASIS to previous state-of-the-art methods. For a fair comparison to the baseline SPADE, we additionally train this model without the feature matching loss and using EMA (Yaz et al., 2018) at the test phase. We refer to this improved baseline as SPADE+.

Synthesis quality. Table 1 compares the image synthesis quality achieved by OASIS and previous methods. In this table, we report the results of our evaluation for OASIS and SPADE+, and the officially reported numbers for all the other models. As seen from Table 1, OASIS outperforms prior state-of-the-art models in FID on all benchmark datasets. Our model also achieves the highest mIoU scores on three out of four datasets, being almost on par with the highest score on ADE20K achieved by SESAME (Ntavelis et al., 2020) Importantly, OASIS achieves the improvement using only adversarial supervision from its segmentation-based discriminator. On the contrary, in the absence of the VGG loss, the baseline SPADE+ does not produce images of high visual quality (see Fig. 1), with two-digit drops in FID scores observed for all the datasets in Table 1. The strong adversarial supervision also allows OASIS to produce images with color and texture distributions closer to the real data. Such improvement over SPADE+ on the ADE20K dataset is shown in Fig. 8, where OASIS achieves the lowest color and texture distances to the target distribution. In contrast, SPADE+ needs to compensate a weaker discriminator signal with the VGG loss, struggling to learn the color and texture distribution of real images without it (see Fig. 8).

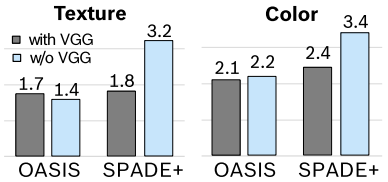

Fig. 8 Histogram distances to real data on the ADE20K validation set. While SPADE+ relies on the VGG loss to learn colors and textures, OASIS achieves low scores without it

Figure 6 shows a qualitative comparison of our results to previous models. Our approach noticeably improves image quality, synthesizing finer textures and more natural colors. While the previous methods occasionally produce areas with unnatural checkerboard artifacts, OASIS generates large objects and surfaces with higher photorealism. Notably, the improvement over previous models is especially remarkable for the semantic classes that occupy large areas, e.g, wall (rows 1,4 in Fig. 6), road (rows 5,6) or water (row 3).

Table 2 Multi-modal synthesis evaluation on ADE20K

Method	Multi-mod.	VGG	MS-SSIM↓	LPIPS↑	FID↓	mIoU↑
SPADE+	Encoder	✓	0.85	0.16	33.4	40.2
SPADE+	3D noise	✗	**0.35**	**0.50**	*58.4*	*18.7*
		✓	0.53	0.36	34.4	36.2
OASIS	3D noise	✗	0.65	0.35	**28.3**	48.8
		✓	*0.88*	*0.15*	31.6	**50.8**

Bold and italic denote the best and the worst performance

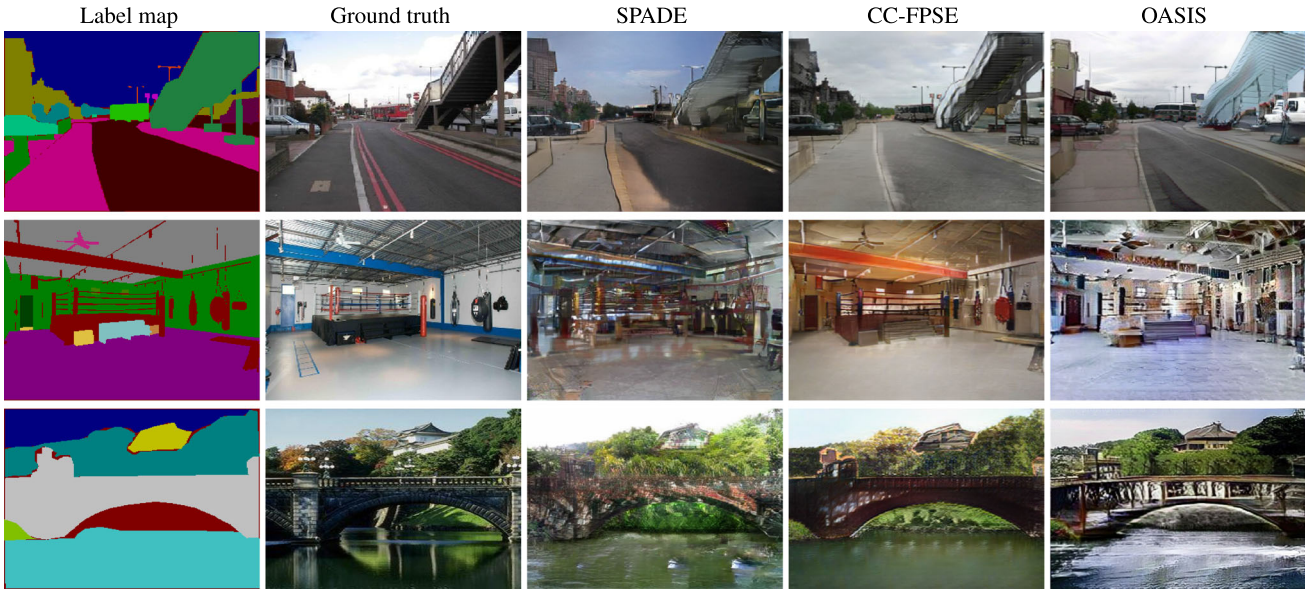

Fig. 9 Failure mode of OASIS. Without the VGG loss, OASIS has less constraints on the diversity in colors and textures. This helps to achieve higher diversity among the generated samples, but sometimes leads to synthesis of objects with outlier colors and textures which may look less realistic compared to Park et al. (2019b) and Liu et al. (2019)

Synthesis diversity. By resampling the input 3D noise, OASIS can produce diverse images given the same label map (see Fig. 2). To measure the diversity of such multi-modal synthesis, we evaluate MS-SSIM (Wang et al., 2003) and LPIPS (Zhang et al., 2018b). The lower the MS-SSIM and the higher the LPIPS scores, the more diverse the generated images are. As seen from Table 2, OASIS outperforms SPADE+ in both diversity metrics, improving the MS-SSIM scores from 0.85 to 0.65 and LPIPS from 0.16 to 0.35. To assess the effect of the perceptual loss and the noise sampling on diversity, we train SPADE+ with 3D noise or the image encoder, and with or without the perceptual loss. Table 2 shows that OASIS, without the perceptual VGG loss, improves over SPADE+ with the image encoder, both in terms of image diversity (MS-SSIM, LPIPS) and quality (mean FID, mIoU across 20 realizations). Using 3D noise further increases diversity for SPADE+. However, a strong quality-diversity trade-off exists for SPADE+: 3D noise improves diversity at the cost of quality, and the perceptual loss improves quality at the cost of diversity. We conclude that our 3D noise injection strongly improves the synthesis diversity, while the VGG loss decreases it.

While the increased diversity is a big advantage, it can also lead to failures in rare cases: for some samples the colors and textures of objects may lie further from the real distribution and seem unnatural to the human eye (see Fig. 9).

4.3 Synthesis Performance on Underrepresented Classes

Class imbalance is a well-known challenge in semantic segmentation applications (Sudre et al., 2017). Similarly to semantic segmentation, to ensure good performance in real-life test scenarios, semantic image synthesis models should account for a possible dataset class imbalance, especially considering that GANs are notorious for dropping modes of training data (Arjovsky & Bottou, 2017). However, to the best of our knowledge, this issue was not addressed in prior works. Thus, in what follows, we evaluate the performance of OASIS and SPADE+ on the ADE20K and LVIS datasets,

Table 3 Per-class IoU scores on ADE20k, grouped by pixel-wise frequency (the fraction of all pixels in the datasets belonging to one class)

Classes IDs	Pixel-wise frequency (%)	mIoU		
		SPADE+	OASIS (w/o α_c)	OASIS (w. α_c)
0–29	86.4	63.7	**69.1**	68.8
30–59	7.2	47.4	52.4	**56.6**
60–89	3.5	45.3	47.0	**51.5**
90–119	1.8	29.3	36.2	**41.5**
120–149	1.0	26.2	31.2	**39.7**
0–149 (all classes)	100	42.4	47.2	**51.6**

Bold denotes the best performance. Training with per-class loss balancing is denoted by α_c

Table 4 Comparison of SPADE+ and OASIS on the LVIS dataset with 1203 classes and a long tail of underrepresented classes

Method	FID ↓	mAP, % ↑	Classes with AP > 0 ↑
SPADE+	26.8	4.56	439
OASIS	**15.3**	**5.38**	**510**
real data	0	6.70	624

Bold denotes the best performance. Last row shows the scores for the LVIS validation set

considering their class imbalances. While the class imbalance in ADE20K is notable (e.g., 86.4% of all image pixels belongs to the 30 best represented classes), this issue is much more amplified in LVIS, which has a long tail of underrepresented classes (see Fig. 7).

Evaluation on ADE20K. OASIS significantly outperforms the SPADE+ baseline in the alignment between generated images and label maps, as measured by mIoU (see Table 1). As shown in Table 3, the improvement in mIoU on ADE20K comes mainly from the better IoU scores achieved for underrepresented semantic classes.

To illustrate this, the semantic classes are sorted by their pixel-wise frequency in the training images, obtained by dividing the number of pixels a class occupies in the dataset by the total number of pixels of all images (2nd column in Table 3). Table 3 highlights that the relative gain in mIoU is especially high for the groups of underrepresented semantic classes, that cover less than 3% of all pixels in the dataset. For these classes, the relative gain over the SPADE+ baseline exceeds 40%. Remarkably, the gain for this group mainly comes from the per-class balancing applied in the OASIS loss function (columns "w/o α_c" and "w. α_c"), which draws the attention of the discriminator to underrepresented semantic classes, thus allowing a higher quality of their generation. This class balancing computes a weight α_c for the losses of each class c on a per-batch basis, for which the total number of pixels in a given batch is divided by the number of pixels belonging to the class (see Eqs. 2 and 3). We note that the possibility to introduce the pixel-wise frequency based balancing requires the loss to be computed separately for each image pixel. This is a unique property of the OASIS discriminator, in contrast to conventional classification-based discriminators, which have to evaluate realism with a single score for images containing both well- and underrepresented classes together.

Evaluation on LVIS. A quantitative comparison between the models on the LVIS dataset is shown in Table 4. In this more extremely imbalanced data regime, the gain of our model is pronounced: OASIS outperforms SPADE+ by a large margin, lowering the FID by 43% (from 26.8 to 15.3). Figure 10 shows a qualitative comparison between the models. OASIS produces images of higher visual quality with more natural colors and textures. In Table 4 we report the mean Average Precision (mAP) of the instance segmentation network evaluated on the set of generated images. OASIS outperforms SPADE+ in mAP by a notable margin (5.38 vs 4.56), thus producing objects with a more realistic appearance and largely reducing the gap to real data (mAP of 6.70). To evaluate the ability of the models to generate underrepresented classes at the tail of the LVIS data distribution, we count the number of classes for which a non-zero AP score is achieved. Table 4 shows that OASIS can model more semantic classes: OASIS achieves a positive AP for 510 semantic classes compared to 439 for SPADE+, thus exhibiting a better capability to synthesize underrepresented classes.

In addition to better handling the class imbalance, OASIS also visually outperforms SPADE+ on the LVIS label maps with a very large proportion of the background class. As seen in Fig. 10 (four rightmost columns), from such label maps, SPADE+ fails to produce plausible images and suffers from mode collapse. In contrast, OASIS successfully deals with such kinds of inputs, producing diverse and visually plausible images even for the least annotated label maps, with the highest proportion of the background class.

In conclusion, we consider long-tailed datasets, such as LVIS, an interesting direction for future work, as the improved synthesis of multiple tail classes under severe imbalance can significantly boost the applicability of semantic image synthesis to real-world applications.

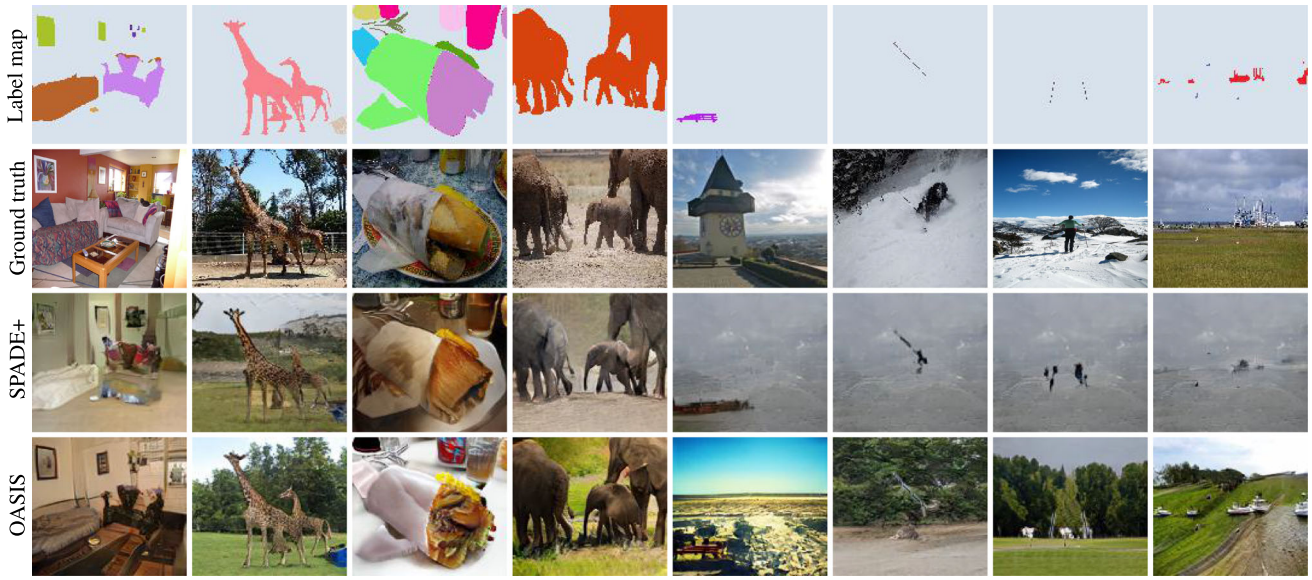

Fig. 10 Qualitative comparison between OASIS and SPADE+ on the long-tailed LVIS dataset with 1203 classes. OASIS generates higher-quality images with more natural colors and textures. For label maps covered mostly by the background class (four right columns), OASIS hallucinates plausible and diverse images, while SPADE+ suffers from mode collapse

Fig. 11 Images generated by OASIS on ADE20K with 256 × 256 resolution using different 3D noise inputs. For both input label maps, the noise is re-sampled globally (first row) or locally in the areas marked in red (second row)

4.4 Image Editing with OASIS

OASIS can generate many different-looking images for a single label map by directly resampling input 3D noise. In the following, we present qualitative multi-modal results and discuss two unique semantic image editing techniques enabled by our model: local resampling of selected semantic classes and diverse resampling of unlabelled images.

Global and local resampling of the 3D noise. The 3D noise of OASIS modulates the activations directly at every

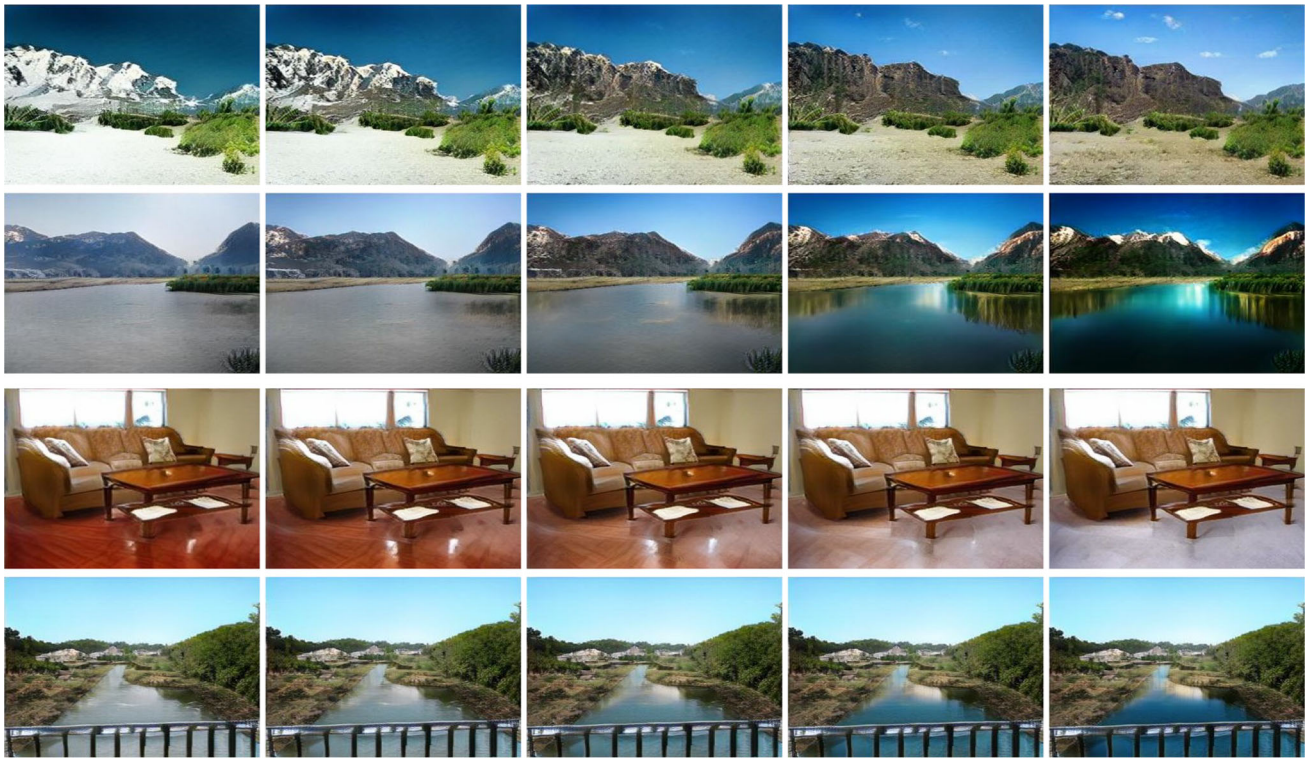

Fig. 12 Latent space interpolations between images generated by OASIS for the ADE20K dataset at resolution 256 × 256. The first two rows display *global* interpolations. The second two rows show *local* interpolations of the floor or water only

generator layer, matching the spatial resolution of features at different generation scales. Therefore, such modulation affects both global and local characteristics of a generated image. At test time, this allows different strategies for noise sampling. For example, the noise can be sampled globally for all pixels, varying the whole image (see Fig. 11, first and third rows). Alternatively, a noise vector can be re-sampled only for specified image regions, resulting in local image editing while preserving the rest of the scene. For example, the local strategy allows to re-sample only the sky area in a landscape scenery, or only the window in a scene of a bedroom (see Fig. 11, second and fourth rows). Spatial sensitivity of OASIS to 3D noise is further demonstrated in Fig. 12, showing interpolations in the latent space. The learned latent space captures well the semantic meaning of objects and allows smooth interpolations not only globally, but also locally for selected objects (see Fig. 12, two last rows).

Creating diverse images from unlabelled data. In contrast to previous semantic image synthesis methods, the OASIS discriminator can be reused as a stand-alone image segmenter. To obtain a segmentation prediction for a given image, a user just needs to feed it to our pre-trained discriminator and select the highest activation among real classes in its $(N + 1)$-channel output for each pixel. When tested as an image segmenter on the validation set of ADE20K, the OASIS discriminator reaches a mIoU of 40.0. For comparison, the state-of-the-art model DeepLab-V3 with a ResNeST backbone (Zhang et al., 2020) achieves an mIoU of 46.91. The good segmentation performance allows OASIS to be applied to unlabelled images: given an unseen image without the ground truth annotation, OASIS can predict a label map via the discriminator. Subsequently feeding this prediction to the generator allows to synthesize a scene with the same layout but different style (see Fig. 13). The recreated scenes closely follow the ground truth label map of the original image and vary considerably, due to the high sensitivity of OASIS to the 3D noise. We note that OASIS uniquely reaches this ability using only adversarial training, without the need for an external segmentation network or additional loss functions. We believe that the ability to create multiple versions of one image while retaining the layout, but not requiring the ground truth label map, may provide useful data augmentation for various applications in future research.

4.5 Synthetic Data Augmentation

As an additional evaluation method, we test the efficacy of generated images when applied as synthetic data augmentation for the task of semantic segmentation. Synthetic data augmentation is a task that benefits from both image quality and diversity, as well as the ability to generate semantic classes that are underrepresented in the original data (see

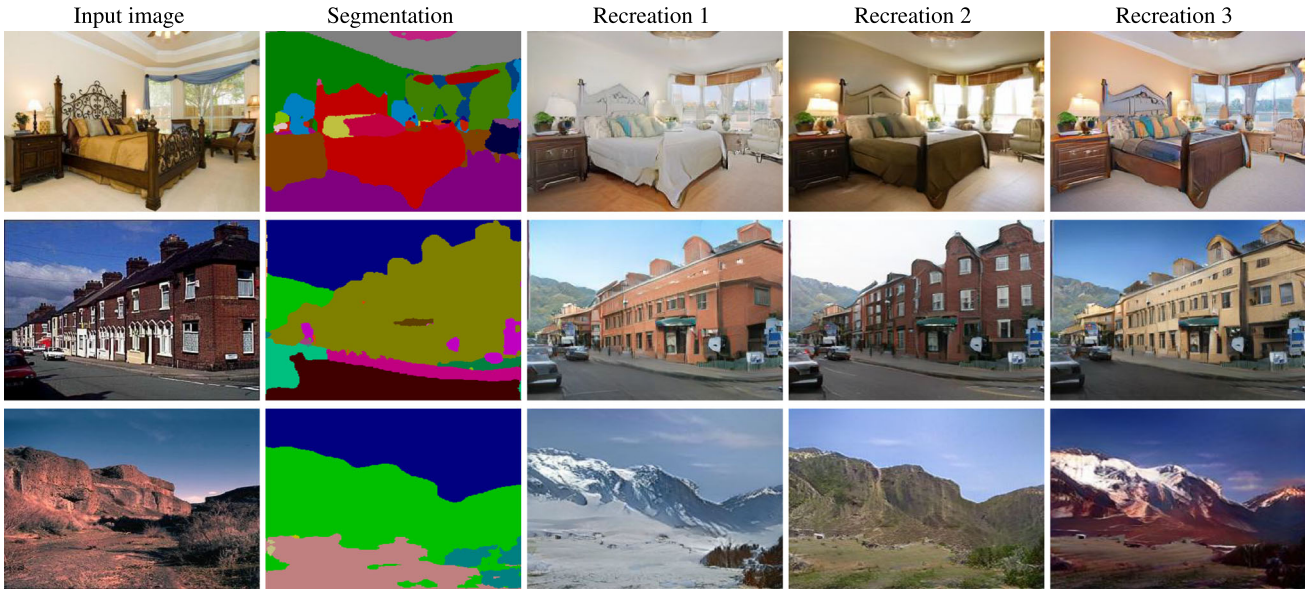

Fig. 13 After training, the OASIS discriminator can be used to segment images. The first two columns show the real image and the segmentation of the discriminator. Using the predicted label map, the generator can produce multiple versions of the original image by resampling noise (Recreations 1–3). Note that no ground truth maps are required

Table 5 Semantic segmentation performance of ResNeSt-50 with and without synthetic data augmentation (DA)

Data augmentation	Cityscapes mIoU↑	ADE20K mIoU↑
No synthetic DA	62.7	41.0
With SPADE	62.6	41.6
With OASIS	**64.7**	**41.8**

Bold denotes the best performance

Table 3). Therefore, the effect of synthetic data augmentation on downstream performance can constitute a more holistic evaluation of semantic image synthesis models. To test the efficiency of OASIS, we train a DeepLab-V3 segmentation network on ADE20K and Cityscapes, at each step augmenting each training image with its synthetic augmentation, produced by OASIS from the same label map.

We compare OASIS against the strong baseline SPADE in Table 5. Between the two methods, OASIS elicits a stronger increase in segmentation performance with an improvement of 2.0 mIoU on Cityscapes and 0.8 mIoU on ADE20K, compared to DeepLab-V3 trained without synthetic augmentation. The higher performance improvement of OASIS compared to SPADE is explained by all the previously observed gains in image quality, diversity, and the alignment to input label maps (see Fig. 8, Tables 1 and 2). In addition to that, the segmentation performance is also improved due to the fact that OASIS tends to synthesize underrepresented classes better than SPADE, which is evident from Table 6. This table compares the IoU performance of DeepLab-V3 on the well represented and underrepresented classes of Cityscapes, as measured by the pixel-wise frequency of the semantic class in the dataset. Examples of well represented classes are road and building (see the 1st row of Table 6), while classes like bicycle or traffic light are the least represented in the dataset (see 4th row in Table 6). Note that the IoU comparison in Table 6 is different from Table 3, where the IoU was measured directly on synthetic data using a pretrained segmenter. It can be seen that the improvement in IoU through OASIS can be mostly attributed to better performance on underrepresented classes, as the gap in performance between OASIS and SPADE becomes larger for the classes which are less represented. Lastly, since the OASIS generator was trained to fool an image segmenter (the OASIS discriminator), it may synthesize harder examples for semantic segmentation than SPADE, thus having higher potential to improve the generalization of segmentation networks to challenging corner cases. We find the above results promising for future utilization of OASIS in various downstream applications. Moreover, for future research, we find it interesting to explore synthetic data augmentation in combination with other data augmentation techniques, e.g., RandAugment (Cubuk et al., 2020), which has the potential to provide further performance gains for downstream applications.

Table 6 Per-class IoU scores on Cityscapes, obtained without (None) and with synthetic data augmentation using SPADE or OASIS

Sorted classes	Pixel-wise frequency (%)	None	SPADE		OASIS	
			abs	rel	abs	rel
0–4	82.7	90.6	90.6	+0.0	**90.9**	**+ 0.3**
5–8	12.5	66.2	66.2	+0.0	**67.4**	**+ 1.2**
9–12	3.3	50.2	49.1	− 1.1	**52.2**	**+ 2.0**
13–18	1.6	51.9	52.3	+0.4	**55.4**	**+3.5**
All classes	100	62.7	62.6	− 0.1	**64.7**	**+ 2.0**

The classes are sorted and grouped by class pixel-wise frequency, as measured by the total fraction of pixels in the dataset belonging to one class. Bold denotes the best performance. The absolute (abs) and relative (rel) mIoU gain via data augmentation is shown

Table 7 Main ablation on ADE20K. The OASIS generator is a lighter version of the SPADE+ generator (72M vs 96M parameters)

G	*D*	VGG	LabelMix	FID↓	mIoU↑
SPADE+	SPADE+	✗	✗	60.7	21.0
SPADE+	OASIS	✗	✗	29.0	**52.1**
OASIS	OASIS	✗	✗	29.3	51.6
		✗	✓	28.4	50.6
OASIS +3D noise	OASIS	✗	✓	**28.3**	48.8
		✓	✓	31.6	50.8

Bold denotes the best performance

Table 8 Ablation on the *D* architecture

D architecture	w/o VGG		with VGG	
	FID↓	mIoU↑	FID↓	mIoU↑
MS-PatchGAN (2x)	60.7	21.0	32.9	42.5
PatchGAN	*197*	*0.62*	34.2	42.2
ResNet-PatchGAN	*147*	*0.42*	32.4	45.1
OASIS	**29.3**	**51.6**	**29.2**	**51.1**

Bold denotes the best performance, italics shows collapsed runs

Table 9 Ablation on the label map encoding runs

Label encoding	w/o VGG		with VGG	
	FID↓	mIoU↑	FID↓	mIoU↑
Input concatenation	*280*	*0.02*	30.0	43.9
Projection	32.4	44.9	**28.0**	46.9
N+1 loss	**28.3**	47.2	28.6	49.8
Balanced N+1 loss	29.3	**51.6**	29.2	**51.1**

Bold denotes the best performance, italics shows collapsed runs

4.6 Ablations

We conduct all our ablations on the ADE20K dataset. We choose this dataset as it more challenging (with 150 classes) than Cityscapes (35 classes) and ADE20K-Outdoors (110 classes), and has more reasonable training time (5 days) compared to COCO-Stuff and LVIS (4 weeks). Our main ablation shows the impact of the main technical components of OASIS, including the new discriminator, lighter generator, LabelMix and the 3D noise. Further ablations are concerned with the architecture changes in the discriminator, the label map encoding in the discriminator, different noise sampling strategies, LabelMix and the GAN feature matching loss.

Main ablation. Table 7 shows that SPADE+ achieves low performance on the image quality metrics without the perceptual loss. Replacing the SPADE+ discriminator with the OASIS discriminator, while keeping the generator fixed, improves FID and mIoU by more than 30 points. Changing the SPADE+ generator to the lighter OASIS generator leads to a negligible degradation of 0.3 in FID and 0.5 in mIoU, but reduces the number of parameters from 96M to 72M. With LabelMix FID improves further by about 1 point. Adding 3D noise improves FID but degrades mIoU, as diversity complicates the task of the pre-trained semantic segmentation network used to compute the mIoU score. For OASIS the perceptual loss deteriorates FID by more than 2 points, but improves mIoU. Overall, without the VGG loss the new discriminator is the key to the performance boost over SPADE+.

Ablation on the discriminator architecture. We train the OASIS generator with three alternative discriminators: the original multi-scale PatchGAN consisting of two networks, a single-scale PatchGAN, and a ResNet-based discriminator, corresponding to the encoder of the U-Net shaped OASIS discriminator. Table 8 shows that the alternative discriminators only perform well with perceptual supervision, while the OASIS discriminator achieves superior performance independent of it. The single-scale discriminators even collapse without the perceptual loss (italic in Table 8).

Table 10 Different 3D noise sampling strategies during training. Bold denotes the best performance

Sampling	Cityscapes			ADE20K		
	FID↓	mIoU↑	MS-SSIM↓	FID↓	mIoU↑	MS-SSIM↓
Image-level	47.7	69.3	0.64	**28.3**	**48.8**	0.65
Region-level	48.1	69.7	**0.62**	28.8	48.1	**0.58**
Pixel-level	50.9	65.5	0.84	28.6	34.0	0.68
Mix	**46.4**	**70.9**	0.68	28.5	47.6	0.66

Ablation on the discriminator label map encoding. We study four different ways to use label maps in the discriminator: the first encoding is input concatenation, as in SPADE. The second option is a pixel-wise projection-based GAN loss (Miyato & Koyama, 2018). Unlike Miyato and Koyama (2018), we condition the GAN loss on the label map instead of a single label. The third and fourth option is to employ the label maps as ground truth for the $N+1$ segmentation loss, or for the class-balanced $N+1$ loss (see Sect. 3.2). For a fair comparison we use neither 3D noise nor LabelMix. As shown in Table 9, input concatenation is not sufficient without additional perceptual loss supervision, leading to training collapse. Without the perceptual loss, the $N+1$ loss outperforms the input concatenation and the projection in both the FID and mIoU metrics. Finally, the class balancing enables enhanced supervision for underrepresented semantic classes, which noticeably improves mIoU scores. On the other hand, we observed that the FID metric is more sensitive to the synthesis of well represented classes and not underrepresented classes, which explains the negative effect of the class balancing on FID.

Ablation on noise sampling strategies for training. Our 3D noise can contain the same sampled vector for each pixel, or different vectors for different regions. This allows for different sampling strategies during training. Table 10 shows the effect of using different methods of sampling 3D noise for different locations during training: *Image-level* sampling creates one global 1D noise vector and replicates it along the height and width of the label map to create a 3D noise tensor. *Region-level* sampling relies on generating one 1D noise vector per semantic class, and stacking them in 3D to match the height and width of the semantic label map. *Pixel-level* sampling creates different noise for every spatial position, with no replication taking place. *Mix* switches between image-level and region-level sampling via a coin flip decision at every training step. With no obvious winner in performance, we choose the simplest scheme (image-level) for our experiments. We find a further investigation with more advanced strategies an interesting direction for future work.

Ablation on LabelMix. Consistency regularization for the segmentation output of the discriminator requires a method of generating binary masks. Therefore, we compare the effectiveness of CutMix (Yun et al., 2019) and our proposed LabelMix. Both methods produce binary masks, but only LabelMix respects the boundaries between semantic classes in the label map. Table 11 compares the FID and mIoU scores of OASIS trained with both methods on the Cityscapes dataset. As seen from the table, LabelMix improves both FID (51.5 vs. 47.7) and mIoU (66.3 vs. 69.3), in comparison to OASIS without consistency regularization. CutMix-based consistency regularization only improves the mIoU (66.3 vs. 67.4), but not as much as LabelMix (69.3). We suspect that since the images are already partitioned through the label map, an additional partition through CutMix results in a dense patchwork of areas that differ by semantic class and real/fake class identity. This may introduce additional label noise during training for the discriminator. To avoid such inconsistency between semantic classes and real/fake identity, the mask of LabelMix is generated according to the label map, providing natural borders between semantic regions, so that the real and fake objects are placed side-by-side without interfering with each other. Under LabelMix regularization, the generator is encouraged to respect the natural semantic class boundaries, improving pixel-level realism while also considering the class segment shapes.

Table 11 Ablation study on the impact of LabelMix and CutMix for consistency regularization (CR) in OASIS on Cityscapes

Transformation	FID↓	mIoU ↑
No CR	51.5	66.3
CutMix	52.1	67.4
LabelMix	**47.7**	**69.3**

Bold denotes the best performance

Ablation on the feature matching loss. We measure the effect of the discriminator feature matching loss (FM) in the absence and presence of the perceptual loss (VGG). The discriminator feature matching loss is used by default in SPADE. Table 12 presents the results for OASIS and SPADE+ on Cityscapes. For SPADE+, we observe that the feature matching loss affects the metrics notably only when no perceptual loss is used. In this case, the FM loss improves mIoU by 8.2 points. In contrast, the effect of the FM loss on the mIoU is small when the perceptual loss is used (0.4 points). Hence, the role of the FM loss in the training of SPADE+ is to improve performance by stabilizing the training, similar to the perceptual loss. This observation is in line with the

Table 12 The effect of the discriminator feature matching loss (FM) in the absence or presence of the perceptual loss (VGG)

VGG	FM	FID↓	mIoU↑
(a) OASIS on Cityscapes			
✗	✗	47.7	69.3
✗	✓	48.5	69.1
✓	✗	**46.1**	**72.0**
✓	✓	46.5	70.9
(b) SPADE+ on Cityscapes			
✗	✗	61.4	47.6
✗	✓	57.3	55.8
✓	✗	**47.8**	64.0
✓	✓	48.1	**64.4**

Bold denotes the best performance

general observation that SPADE and other semantic image synthesis models require the help of additional loss functions because the adversarial supervision through the discriminator is not strong enough. In comparison, we did not observe any training collapses in OASIS, despite not using any extra loss functions. For OASIS, the feature matching loss results in a worse FID (by 0.8 points) in the absence of the perceptual loss. We also observe a degradation of 1.1 mIoU points through the FM loss, in the case where the perceptual supervision is present. This indicates that the FM loss negatively affects the strong supervision from the semantic segmentation adversarial loss of OASIS.

5 Conclusion

This work studies semantic image synthesis, the task of generating diverse and photorealistic images from semantic label maps. Conventionally, semantic image synthesis GAN models employed a perceptual VGG loss to overcome training instabilities and improve the synthesis quality. In our experiments we demonstrated that the VGG-based perceptual loss imposes unnecessary constraints on the feature space of the generator, significantly limiting its ability to produce diverse samples from input noise, as well as the ability to produce images with colors and textures closely matching the distribution of real images. Therefore, in this work we propose OASIS, a semantic image synthesis model that needs only adversarial supervision to achieve high-quality results.

The improvement over the prior work in image synthesis quality is achieved via the detailed spatial and semantic-aware supervision from our novel segmentation-based discriminator, which uses semantic label maps as ground truth for training. With this powerful discriminator, OASIS can easily generate diverse outputs from the same semantic label map by resampling 3D noise, eliminating the need for additional image encoders to achieve multi-modality. The proposed 3D noise injection scheme can work both in a global and local regime, allowing to change the appearance of the whole scene and of individual objects. With the proposed modifications, OASIS significantly improves over previous state-of-the-art models in terms of image synthesis quality.

Furthermore, we proposed to use the LVIS dataset to evaluate semantic image synthesis under severe class imbalance and sparse label annotations. Thanks to the class balancing mechanism enabled by its segmentation-based discriminator, OASIS achieves more realistic synthesis of underrepresented classes, achieving pronounced gains on the extremely unbalanced LVIS dataset. Lastly, the design of OASIS can be better suited for image editing applications compared to the SPADE baseline, enabling diverse resampling of scenes from unlabelled images, as well as for synthetic data augmentation, improving the performance of a downstream segmentation network by a larger margin.

Acknowledgements Juergen Gall has been 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

Alharbi, Y., & Wonka, P. (2020). Disentangled image generation through structured noise injection. In *Conference on computer vision and pattern recognition (CVPR)*.

Arjovsky, M., & Bottou, L. (2017). Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations (ICLR)*.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2016). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *Transactions on Pattern Analysis and Machine Intelligence, 39*, 2481–2495.

Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *International conference on learning representations (ICLR)*.

Bruna, J., Sprechmann, P., & LeCun, Y. (2016). Super-resolution with deep convolutional sufficient statistics. In *International conference on learning representations (ICLR)*.

Caesar, H., Uijlings, J., & Ferrari, V. (2018). Coco-stuff: Thing and stuff classes in context. In *Conference on computer vision and pattern recognition (CVPR)*.

Casanova, A., Careil, M., Verbeek, J., Drozdzal, M., & Romero Soriano, A. (2021). Instance-conditioned gan. In *Advances in neural information processing systems (NeurIPS)*.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2015). Semantic image segmentation with deep convolutional nets and fully connected crfs. In *International conference on learning representations (ICLR)*.

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision (ECCV)*.

Chen, Q., & Koltun, V. (2017). Photographic image synthesis with cascaded refinement networks. In *International conference on computer vision (ICCV)*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Conference on computer vision and pattern recognition (CVPR)*.

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in neural information processing systems (NeurIPS)*.

Deng, J., Dong, W., Socher, R., Li, LJ., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Conference on computer vision and pattern recognition (CVPR)*.

Gatys, L., Ecker, A. S., Bethge, M. (2015). Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems (NeurIPs)*.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Conference on computer vision and pattern recognition (CVPR)*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems (NeurIPS)*.

Gupta, A., Dollar, P., & Girshick, R. (2019). LVIS: A dataset for large vocabulary instance segmentation. In *Conference on computer vision and pattern recognition (CVPR)*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in neural information processing systems (NeurIPS)*.

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *European conference on computer vision (ECCV)*.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Conference on computer vision and pattern recognition (CVPR)*.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision (ECCV)*.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Conference on computer vision and pattern recognition (CVPR)*.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020a). Training generative adversarial networks with limited data. In *Advances in neural information processing systems (NeurIPS)*.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020b). Analyzing and improving the image quality of stylegan. In *Conference on computer vision and pattern recognition (CVPR)*.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. In *Advances in neural information processing systems (NeurIPS)*.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations (ICLR)*.

Li, K., & Malik, J. (2018). Implicit maximum likelihood estimation. arXiv:1809.09087.

Li, K., Zhang, T., & Malik, J. (2019). Diverse image synthesis from semantic layouts via conditional imle. In *International conference on computer vision (ICCV)*.

Li, Y., Li, Y., Lu, J., Shechtman, E., Lee, Y. J., & Singh, K. K. (2021). Collaging class-specific gans for semantic image synthesis. In *International conference on computer vision (ICCV)*.

Liu, B., Zhu, Y., Song, K., & Elgammal, A. (2021). Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International conference on learning representations (ICLR)*.

Liu, X., Yin, G., Shao, J., Wang, X., & Li, H. (2019). Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in neural information processing systems (NeurIPS)*.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv:1411.1784

Miyato, T., & Koyama, M. (2018). cGANs with projection discriminator. In *International conference on learning representations (ICLR)*.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In *International conference on learning representations (ICLR)*.

Ntavelis, E., Romero, A., Kastanis, I., Van Gool, L., & Timofte, R. (2020). Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. In *European conference on computer vision (ECCV)*.

Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition, 29*, 51–59.

Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019a). Gaugan: Semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH*.

Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019b). Semantic image synthesis with spatially-adaptive normalization. In *Conference on computer vision and pattern recognition (CVPR)*.

Park, T., Efros, A. A., Zhang, R., & Zhu, J. Y. (2020). Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision (ECCV)*.

Qi, X., Chen, Q., Jia, J., & Koltun, V. (2018). Semi-parametric image synthesis. In *Conference on computer vision and pattern recognition (CVPR)*.

Reed, S. E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. In *International conference on machine learning (ICML)*.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., & Cohen-Or, D. (2021). Encoding in style: A stylegan encoder for image-to-image translation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision (IJCV), 40*, 99–121.

Sauer, A., Chitta, K., Müller, J., & Geiger, A. (2021). Projected gans converge faster. In *Advances in neural information processing systems (NeurIPS)*.

Schönfeld, E., Schiele, B., & Khoreva, A. (2020). A u-net based discriminator for generative adversarial networks. In *Conference on computer vision and pattern recognition (CVPR)*.

Schönfeld, E., Sushko, V., Zhang, D., Gall, J., Schiele, B., & Khoreva, A. (2021). You only need adversarial supervision for semantic image synthesis. In *International conference on learning representations (ICLR)*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations (ICLR)*.

Souly, N., Spampinato, C., & Shah, M. (2017). Semi supervised semantic segmentation using generative adversarial network. In *International conference on computer vision (ICCV)*.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*.

Tan, Z., Chen, D., Chu, Q., Chai, M., Liao, J., He, M., Yuan, L., & Yu, N. (2020). Rethinking spatially-adaptive normalization. arXiv:2004.02867.

Tang, H., Bai, S., & Sebe, N. (2020a). Dual attention gans for semantic image synthesis. In *ACM international conference on multimedia*.

Tang, H., Qi, X., Xu, D., Torr, P. H., & Sebe, N. (2020b). Edge guided gans with semantic preserving for semantic image synthesis. arXiv:2003.13898.

Tang, H., Xu, D., Yan, Y., Torr, PH. ., & Sebe, N. (2020c). Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *Conference on computer vision and pattern recognition (CVPR)*.

Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C. C., & Lin, D. (2021a). Seesaw loss for long-tailed instance segmentation. In *Conference on computer vision and pattern recognition (CVPR)*.

Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional GANs. In *Conference on computer vision and pattern recognition (CVPR)*.

Wang, Y., Qi, L., Chen, Y. C., Zhang, X., & Jia, J. (2021b). Image synthesis via semantic composition. In *ICCV*.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *Asilomar conference on signals, systems & computers*.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018). Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*.

Yaz, Y., Foo, C. S., Winkler, S., Yap, K. H., Piliouras, G., & Chandrasekhar, V. (2018). The unusual effectiveness of averaging in gan training. In *International conference on learning representations (ICLR)*.

Yu, F., Koltun, V., & Funkhouser, T. (2017) Dilated residual networks. In *Conference on computer vision and pattern recognition (CVPR)*.

Yun, S,, Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International conference on computer vision (ICCV)*.

Zhang, D., & Khoreva, A. (2019). PA-GAN: Improving GAN training by progressive augmentation. In *Advances in neural information processing systems (NeurIPS)*.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2018a). StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 41,* 1947–1962.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., & Smola, A. (2020). Resnest: Split-attention networks. arXiv:2004.08955.

Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., & Yang, Y. (2021). Cross-modal contrastive learning for text-to-image generation. In *Conference on computer vision and pattern recognition (CVPR)*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018b). The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on computer vision and pattern recognition (CVPR)*.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Conference on computer vision and pattern recognition (CVPR)*.

Zhu, Z., Xu, Z., You, A., & Bai, X. (2020). Semantically multi-modal image synthesis. In *Conference on computer vision and pattern recognition (CVPR)*.