Pattern Recognition 166 (2025) 111684

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Dual-function discriminator for semantic image synthesis in variational GANs

Aihua Ke [a], Bo Cai [a,*], Yujie Huang [a], Jian Luo [a], Yaoxiang Yu [a], Le Li [b]

[a] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China

[b] School of Mathematics and Statistics, Central China Normal University, Wuhan, China

ARTICLE INFO

Keywords:
Semantic image synthesis
Generative adversarial networks
Conditional residual attention module
Dual-function discriminator

ABSTRACT

Semantic image synthesis aims to generate target images conditioned on given semantic labels, but existing methods often struggle with maintaining high visual quality and accurate semantic alignment. To address these challenges, we propose VD-GAN, a novel framework that integrates advanced architectural and functional innovations. Our variational generator, built on an enhanced U-Net architecture combining a pre-trained Swin transformer and CNN, captures both global and local semantic features, generating high-quality images. To further boost performance, we design two innovative modules: the Conditional Residual Attention Module (CRAM) for dimensionality reduction modulation and the Channel and Spatial Attention Mechanism (CSAM) for extracting key semantic relationships across channel and spatial dimensions. Additionally, we introduce a dual-function discriminator that not only distinguishes real and synthesized images, but also performs multi-class segmentation on synthesized images, guided by a redefined class-balanced cross-entropy loss to ensure semantic consistency. Extensive experiments show that VD-GAN outperforms the latest supervised methods, with improvements of (FID, mIoU, Acc) by (5.40%, 4.37%, 1.48%) and increases in auxiliary metrics (LPIPS, TOPIQ) by (2.45%, 23.52%). The code will be available at https://github.com/ah-ke/VD-GAN.git.

1. Introduction

Recent advances in deep-generative models have significantly propelled both unconditional and conditional generation. Unconditional generation involves random noise as input, while conditional generation uses additional conditions, such as descriptive text [1], scene images [2,3], or semantic labels [4]. When the condition is a semantic label, the process is known as label-to-image synthesis, focusing on generating high-fidelity, semantically accurate images from segmented labels, as shown in Fig. 1. This technique is useful for applications like content creation, image editing, and data augmentation, where new similar samples can be generated to improve the training of deep learning models.

Methods for semantic image synthesis are typically divided into supervised and unsupervised approaches. Unsupervised methods use unpaired data to generate realistic images from semantic maps, often employing loss of cycle consistency to maintain image relationships [2,5]. Supervised methods, using labeled data, tend to produce higher quality images, focusing on areas such as normalization functions [4], attention mechanisms [6], and discriminator networks [7]. Although diffusion models, compared to GANs, offer good generative performance, they involve complex noise sampling, slower generation, and higher resource consumption. Additionally, aligning synthesized images with semantic labels remains challenging, especially when dealing with imbalanced datasets and large image volumes.

Despite their successes, GAN models still face challenges in image quality and semantic alignment. The key issues that contribute to these challenges include the following. First, many GAN models use CNN-based decoders as generators for image synthesis from semantic labels, but the limitations of CNNs can hinder the extraction of global features and slow down training. Second, earlier GAN models often combine the given semantic layout with 3D Gaussian noise as inputs to the UNet structure, which can lead to suboptimal use of the available input information. Third, in previous residual modules, the output of the pre-layer is directly added to the input of the post-layer via shortcut connections, potentially passing redundant information that increases the network's computational load. Lastly, while existing GAN models prioritize enhancing the photo-realism of synthesized images, they often overlook the importance of maintaining strong semantic

* Corresponding author.
E-mail address: caib@whu.edu.cn (B. Cai).

https://doi.org/10.1016/j.patcog.2025.111684
Received 7 October 2023; Received in revised form 27 March 2025; Accepted 2 April 2025
Available online 11 April 2025

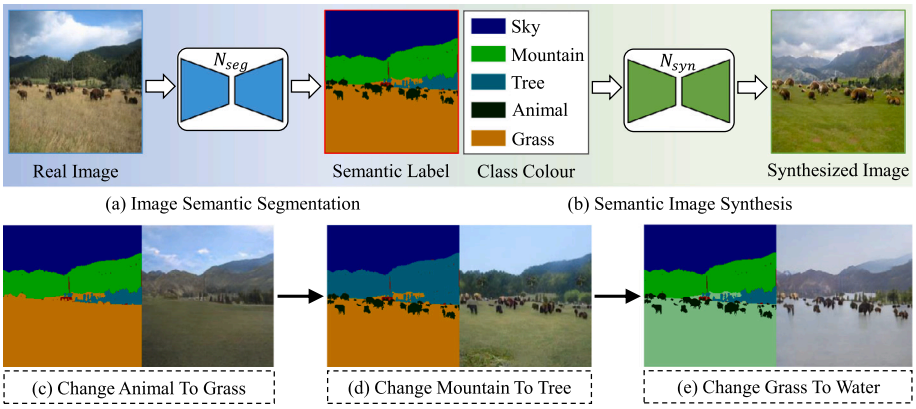


Fig. 1. (a) Image semantic segmentation: The segmentation network N_{seg} generates a semantic label from a real image, marking different classes with different colors. (b) Semantic Image Synthesis: The semantic label serves as the input condition for the synthesis network N_{syn} to generate a high-quality image. (c–e) Example Application: Demonstrates semantic control synthesis using our method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alignment between the generated images and the provided semantic labels.

To address these limitations, we propose a novel supervised GAN-based method, VD-GAN. Our approach introduces a variational generator that enhances image quality by building upon an improved U-Net structure, which integrates both the pre-trained Swin Transformer and CNN. Specifically, CNN excels at capturing local features such as edges and textures, while the Swin Transformer effectively models global semantic information through its self-attention mechanism. By combining these two components, our generator captures both local details and global context simultaneously, resulting in images with improved visual coherence and quality. In addition, we introduce two novel modules to further enhance the generator's performance: the Conditional Residual Attention Module (CRAM) and the Channel and Spatial Attention Mechanism (CSAM). Traditional models often struggle to capture semantic relationships across different channels and spatial dimensions, which can lead to inefficient feature extraction. CSAM addresses this by dynamically focusing on the most relevant channels and spatial regions, thereby improving the capture and utilization of critical semantic information. CRAM refines the reduction of dimensions while preserving essential semantic details. Beyond distinguishing between real and synthesized images, our dual-function discriminator also performs multi-class segmentation. This multi-task feature improves semantic alignment by incorporating a redefined class-balanced cross-entropy loss and a multi-scale strategy, which enhances the discriminator's ability to accurately guide the generator.

In summary, our approach introduces several key innovations: the integration of Swin Transformer and CNN into the U-Net architecture for enhanced hierarchical feature extraction, the Conditional Residual Attention Module and Channel and Spatial Attention Mechanism for refined dimensionality reduction and improved semantic relations, and the dual-function discriminator for better semantic alignment. Extensive experiments on the Cityscapes [8], ADE20K [9], and COCO-Stuff [10] datasets demonstrate that VD-GAN outperforms state-of-the-art methods in both visual quality and semantic alignment.

2. Related work

2.1. Generative adversarial networks

Semantic image synthesis methods primarily rely on Generative Adversarial Networks (GANs) [6], which consist of a generator and a discriminator. The generator produces high-quality images from semantic labels, while the discriminator distinguishes between synthesized and real images. Recent advancements in conditional GANs, such as SPADE [11] and OCGAN [4], have enhanced image generation by incorporating perceptual losses and more complex input mechanisms for better semantic consistency. However, these methods still face challenges, such as more difficult training and limitations in multi-scale semantic extraction.

2.2. Semantic image synthesis

Semantic image synthesis [12] involves generating high-quality images from provided semantic labels, often leveraging GANs. For instance, GauGAN [11] modulates normalization layers with affine transformations based on the input semantic layout, while SAFM [7] introduces a shape-aware position descriptor to represent object shapes and pixel positions. In addition to GAN-based approaches, non-GAN methods, such as variational models (e.g., VAE [13] and VASIS [14]) and newer diffusion-based techniques [15], have also emerged as alternatives. Variational models excel in capturing latent distributions but often suffer from blurriness and semantic misalignment. Despite these challenges, advancements in these models highlight the need for more efficient strategies to better extract and utilize critical semantic information.

2.3. Discriminator network architectures

In GAN models, the discriminator is crucial for guiding the generator's learning. As semantic image synthesis has progressed, discriminator designs have evolved. Pix2Pix [16] introduced PatchGAN, which uses a convolutional structure to classify image patches as real or fake. Pix2PixHD [17] enhanced this with a multi-scale approach, while CC-FPSE [18] added a feature pyramid semantic embedding discriminator using both images and label maps. OASIS [19] introduced a segmentation-based discriminator for better semantic alignment. Despite these improvements, limitations remain, indicating the need for further refinement.

3. Method

3.1. Overall framework

We present VD-GAN, a variational generative adversarial network equipped with a dual-function discriminator (Fig. 2). The model consists of two main components: a variational generator and a dual-function discriminator. The generator leverages a pre-trained Swin transformer combined with CNNs to ensure faster, more stable convergence. It incorporates a Conditional Residual Attention Module (CRAM) in each decoder layer to optimize feature extraction and improve the utilization of input information. The dual-function discriminator performs two key tasks: distinguishing between real and synthesized

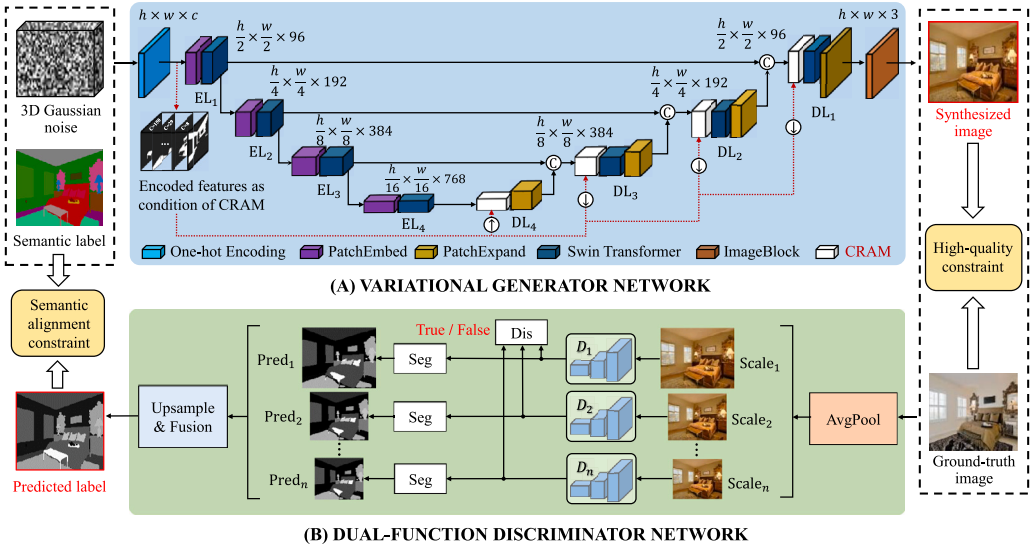


Fig. 2. The overall framework of the proposed VD-GAN method is depicted. (A) The variational generator network employs an enhanced U-Net structure to synthesize high-quality images conditioned on semantic labels, incorporating the Conditional Residual Attention Module (CRAM). (B) The dual-function discriminator network D operates in a multi-scale manner to implement true/false distinguishing and multi-class segmentation prediction on the synthesized image.

images and conducting multi-class segmentation on the generated images. A class-balanced cross-entropy loss reduces the gap between predicted and semantic labels, while a multi-scale strategy further improves performance. VD-GAN is trained end-to-end in a supervised manner through a mini-max game. Further details on both networks are provided in the following sections.

3.2. Variational generator network

As shown in Fig. 2(A), our variational generator G employs an enhanced U-Net architecture [16] to generate high-fidelity images from semantic labels. The U-Net consists of an encoder and a decoder. The encoder extracts hierarchical features from the combined input of the semantic label and 3D Gaussian noise. To address the discrete nature of semantic labels, we apply one-hot encoding [9] to map the input to a continuous latent space. In the encoder (EL_i), we replace traditional convolution and down-sampling with the Swin transformer for improved global feature extraction and dependency modeling. As the network deepens, the encoder captures features with higher dimensionality and lower resolution, preserving more spatial information from the input labels.

The decoder reconstructs the synthesized image by combining features from the encoder layer (EL_i) and the previous decoder layer (DL_{i-1}). These features are processed by the Conditional Residual Attention Module (CRAM) to reduce dimensionality while preserving key information. We replace deconvolution and up-sampling with a Swin transformer block and patch expansion [20], addressing checkerboard artifacts and improving detail recovery. The final output is generated through an ImageBlock operation, consisting of a convolution block followed by a tanh activation. Further details on CRAM are provided in subsequent sections.

3.2.1. Conditional residual attention module

In our U-Net decoder, shallow and deep features are combined along the channel axis to address information loss during encoding. However, this increases computational demands due to high-dimensional features. To reduce this, we introduce the Conditional Residual Attention Module (CRAM), which replaces the standard 1×1 convolution [16]. CRAM preserves spatial and semantic information by reducing dimensionality through a shortcut connection and conditional features. As shown in Fig. 3(a), CRAM consists of two BasicBlock functions with Conditional Batch Normalization (CBN), Leaky ReLU, and a 3×3 convolution. The BasicBlock update is described by Eq. (1):

$$R_{i+1} = \mathcal{F}(R_i), \tag{1}$$

where R_{i+1} is the updated residual feature, and $\mathcal{F}$ is the feature update function. This update is computed as shown in Eq. (2):

$$\mathcal{F}(R_i) = \mathcal{H}(F_s, \delta(\psi(R_i))), \tag{2}$$

where $\mathcal{H}$ represents the CBN operation, δ is the activation function, and ψ denotes the convolution operation. F_s refers to the conditional features that help adapt the CBN based on the semantic input, improving feature extraction efficiency. After two BasicBlock functions, CRAM combines the outputs through element-wise addition:

$$R_{i+2} = \mathcal{F}(R_{i+1}) + \xi(R_i), \tag{3}$$

where $\xi(R_i)$ is the Channel and Spatial Attention Mechanism (CSAM) embedded in CRAM. This mechanism refines important features and suppresses irrelevant ones, ensuring improved semantic representation.

The Conditional Batch Normalization (CBN) layer structure is shown in Fig. 3(b). CBN normalizes input activations h and scales and shifts them with learned parameters γ and β, as described in Eq. (4):

$$\mathcal{H}(F_s, h) = \frac{h_{n,c,x,y} - \mu_c}{\sigma_c} \gamma_{c,x,y}(F_s) + \beta_{c,x,y}(F_s), \tag{4}$$

where μ_c and σ_c are the mean and standard deviation of activations, calculated across spatial dimensions. The parameters γ and β are learned using depth-wise separable convolutions ψ^{dp}, as shown in Eqs. (5) and (6):

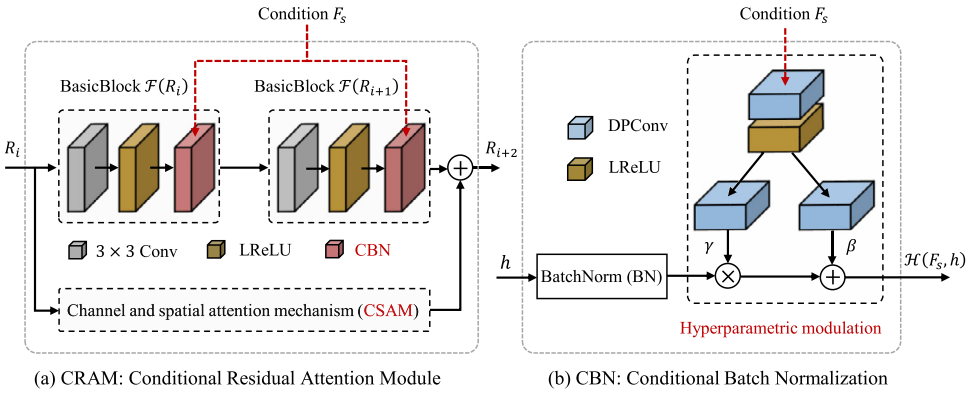
$$\gamma = \psi^{dp}(\delta(\psi^{dp}(F_s))), \tag{5}$$

$$\beta = \psi^{dp}(\delta(\psi^{dp}(F_s))), \tag{6}$$

Overall, CRAM enhances both computational efficiency and semantic alignment in image synthesis, improving photorealism and accuracy. In the next subsection, we will discuss the CSAM structure, which plays a critical role in semantic refinement.

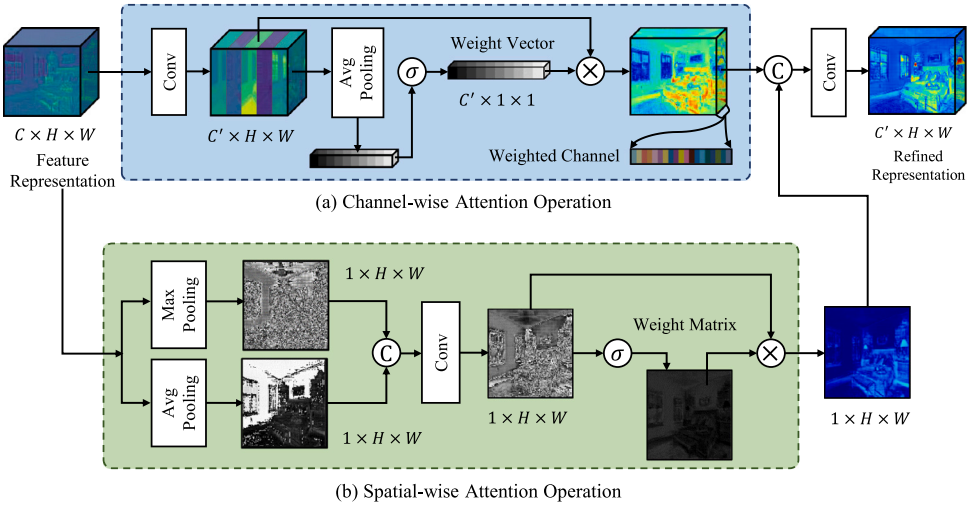
3.2.2. Channel and spatial attention mechanism

In previous residual modules, the pre-layer output is added directly to the post-layer input, leading to redundant information and increased network load. Traditional attention mechanisms [6], such as channel



(a) CRAM: Conditional Residual Attention Module (b) CBN: Conditional Batch Normalization

Fig. 3. (a) The architecture of the Conditional Residual Attention Module (CRAM) integrated into the generator's decoder. CRAM reduces dimensionality while preserving spatial and semantic information. (b) The Conditional Batch Normalization (CBN) layer dynamically modulates activations using two learnable hyperparameters derived from the conditioning input F_s. These parameters scale and shift the feature responses, enabling adaptive feature transformation. The symbols ⊗ and ⊕ denote element-wise multiplication and addition, respectively.



(a) Channel-wise Attention Operation

(b) Spatial-wise Attention Operation

Fig. 4. The structure of the Channel and Spatial Attention Mechanism (CSAM), designed to refine representations with significant semantic and spatial layout information. The CSAM, integrated into the shortcut connection, comprises (a) channel-wise attention operation and (b) spatial-wise attention operation. Symbols ⊗, Ⓒ, and Ⓢ denote element-wise multiplication, channel concatenation, and the sigmoid function, respectively.

or spatial attention, address this by emphasizing specific features, but still allow some redundancy. To solve this, we introduce a Channel and Spatial Attention Mechanism (CSAM) at the shortcut connection (Fig. 4), which more effectively extracts important features before combining them with the post-layer input. CSAM is a lightweight, efficient mechanism that enhances semantic and spatial information.

CSAM consists of two parts: channel-wise and spatial-wise attention. The channel-wise attention (Fig. 4(a)) uses a 1 × 1 convolution, followed by adaptive average pooling and sigmoid activation, to create a weight vector for each channel, emphasizing semantically relevant ones. The spatial-wise attention (Fig. 4(b)) focuses on spatial relationships by combining adaptive average and max pooling, then using 1 × 1 convolution and sigmoid activation to generate a weight matrix for each pixel, highlighting significant spatial areas. CSAM combines these weighted features to produce a refined representation with stronger object boundaries and enhanced feature quality.

3.3. Dual-function discriminator network

In GAN-based semantic image synthesis, the discriminator's goal is to distinguish between real and fake images while ensuring semantic alignment with given labels. To improve semantic alignment, we introduce a multi-class segmentation function within the discriminator. This allows the discriminator to not only assess image authenticity but also perform detailed semantic analysis. Additionally, our multi-scale strategy enhances both discrimination and segmentation performance, improving synthesized image quality and semantic accuracy.

As shown in Fig. 2(B), the dual-function discriminator processes input images at multiple scales using adaptive average pooling. Each scaled image passes through a discriminator with convolutional blocks, instance normalization, and Leaky ReLU activation. The Dis layer outputs a matrix at each scale, with the output at the ith scale defined as Eq. (7):

$$M_i(\mathrm{x}) = \mathcal{F}_i^{\mathrm{dis}}\left(D_i\left(\phi_i(\mathrm{x})\right)\right), \tag{7}$$

where x is the ground-truth image, ϕ_i is the scaling operation, and $\mathcal{F}_i^{\mathrm{dis}}$ applies a convolutional operation. The discriminator uses a hinge-based adversarial loss function to distinguish real and fake images, as shown in Eqs. (8) and (9):

$$\mathcal{L}_{\mathrm{hadv}}^{i} = -\mathbb{E}_x\left[\min\left(-1 + M_i(\mathrm{x}), 0\right)\right] - \mathbb{E}_{(\mathrm{l},\mathrm{z})}\left[\min\left(-1 - M_i(G(\mathrm{l},\mathrm{z})), 0\right)\right] \tag{8}$$

$$\mathcal{L}_{hadv} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{hadv}^{i} \tag{9}$$

where (l, z) is the concatenation of the semantic label and Gaussian noise.

For the segmentation function, the Seg layer generates a multi-class prediction map at each scale, which is up-sampled to the original

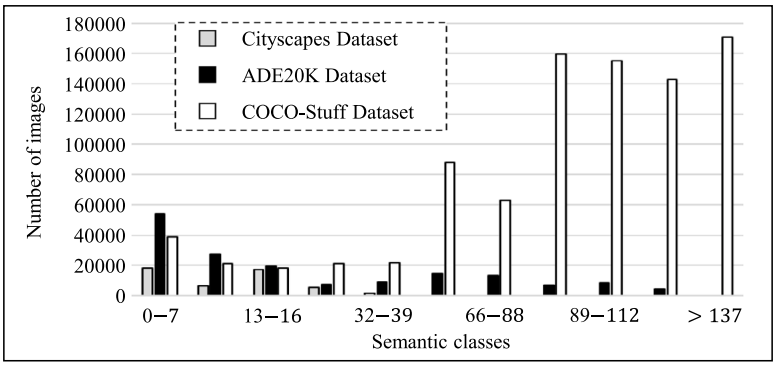


Fig. 5. Distribution in amount of images corresponding to each semantic classes on the public datasets of Cityscapes, ADE20K and COCO-Stuff. The figure highlights the imbalanced nature of semantic class distributions in these datasets.

resolution, as shown in Eq. (10):

$$P(\mathrm{x})_{x,y,c} = \frac{1}{n}\sum_{i=0}^{n} \phi_0\left(\mathcal{F}_i^{\mathrm{seg}}\left(D_i\left(\phi_i(\mathrm{x})\right)\right)\right), \tag{10}$$

Here, ϕ_0 represents the up-sampling operation, and $\mathcal{F}_i^{\mathrm{seg}}$ applies a segmentation layer with up-sampling and a 3×3 convolution. The prediction map includes a 'false' class for synthesized images. The discriminator is trained to align the generated images with semantic labels using a class-balanced cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{ce} = &-\mathbb{E}_{(\mathrm{x},\mathrm{l})}\left[\sum_{c=1}^{N}\alpha_c \sum_{x,y}^{H\times W} \mathrm{l}_{x,y,c}\log P(\mathrm{x})_{x,y,c}\right] \\ &-\mathbb{E}_{(\mathrm{l},\mathrm{z})}\left[\sum_{x,y}^{H\times W}\log P(G(\mathrm{l},\mathrm{z}))_{x,y,c=N+1}\right],\end{aligned} \tag{11}$$

$$\alpha_c = \mathbb{E}_{\mathrm{l}}\left[\frac{H\times W}{\sum_{x,y}^{H\times W}\mathrm{l}_{x,y,c}}\right]. \tag{12}$$

Considering the typical imbalance among the N semantic classes (as shown in Fig. 5), the class balancing weight α_c is calculated as the inverse of the per-pixel class frequency to mitigate the risk of overfitting.

4. Experiments

4.1. Experimental settings

In this paper, the specifics of the experimental settings will be presented in terms of three aspects: datasets, evaluation metrics and implementation details. Further details are as follows:

4.1.1. Datasets

To validate the superiority of the proposed VD-GAN method, we conducted extensive experiments on three public datasets: Cityscapes [8], ADE20K [9], and COCO-Stuff [21]. The Cityscapes dataset contains 2975 training images and 500 validation images, as well as 35 semantic classes. The ADE20K is composed of 20,210 images for training and 2000 images for validation, as well as 150 semantic classes. The COCO-Stuff comprises 118,287 images for training and 5000 images for validation, along with 182 semantic classes. Fig. 5 exhibits the distribution of the number of images corresponding to each semantic class on the cityscape, ADE20K and COCO-Stuff datasets. It is obviously appreciated from the figure that there is an in-homogeneous distribution of semantic classes. In order to verify the robustness of the proposed VD-GAN under different image resolutions, we additionally adjusted the resolutions of the images in the Cityscapes, ADE20K, and COCO-Stuff datasets to 256×512, 256×256, and 256×256, respectively.

4.1.2. Evaluation metric

In this paper, the semantic image synthesis task has two main goals: generating photo-realistic and diverse images, and aligning them well with given semantic labels. Referring to previous work [4], we adopt the Fréchet Inception Distance (FID) [22] as an image generation score to assess the perceptual quality and diversity of the synthesized images. FID compares the feature distributions of real and generated images, with a lower score indicating higher visual fidelity and diversity. Moreover, we also utilize both the mean Intersection over Union (mIoU) and the pixel Accuracy (Acc) as semantic segmentation scores to measure the alignment accuracy of the synthesized image with the semantic labels. mIoU computes the overlap between the predicted and ground truth regions across all classes, where higher values indicate better segmentation alignment. Pixel Accuracy (Acc) measures the percentage of correctly classified pixels in the image, providing an additional metric for the precision of label alignment. To calculate the semantic segmentation scores, we employ the advanced segmentation networks for each experimental dataset: DRN-D-105 for Cityscapes, UperNet101 for ADE20K, and DeepLabV2 [23] for COCO-Stuff.

4.1.3. Implementation details

We utilize the adam optimizer [24] with $\beta_1 = 0$ and $\beta_2 = 0.999$. The learning rates for the generator, lr_g, are set to 0.0002, 0.0004, and 0.0001 for the Cityscapes, ADE20K, and COCO-Stuff datasets, respectively, while the learning rates for the discriminator, lr_d, are set to 0.0002, 0.0004, and 0.0004 for these datasets, respectively. To optimize these learning rates, we use a dynamic adjustment mechanism that linearly decays the learning rate based on the number of training epochs. This mechanism helps stabilize training by gradually reducing the learning rate. Additionally, we follow best practices from related research and use a control variable approach to adjust other parameters according to our specific experimental needs. We train the model for 200 epochs on the Cityscapes and ADE20K datasets to identify the optimal settings, and for 100 epochs on the COCO-Stuff dataset due to its larger number of images. A batch size of 32 is employed across all datasets, and the random seed is set to 1024 to ensure reproducibility. To enhance the model's robustness and evaluation metrics, we apply an exponential moving average [24] to the model parameters with a decay rate of 0.9999. The model is built on the PyTorch deep learning framework, leverages CUDA version 11.6, and is trained in parallel on four NVIDIA Tesla A100 GPUs, each with 40 GB of video memory.

4.2. Quantitative results

Previous semantic image synthesis methods can be broadly classified into two categories: unsupervised and supervised. Accordingly, Tables 1 and 2 present the quantitative comparison results of our method against unsupervised and supervised approaches, respectively, based on image generation scores (FID) and semantic segmentation

Table 1
Quantitative comparison of our method with the unsupervised methods in image generation score (FID) and semantic segmentation score (mIoU) on three public datasets. "↓" denotes the lower the performance the better, while "↑" denotes the higher the performance the better.

Method	Cityscapes		ADE20K		COCO-Stuff	
	FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑
DistanceGAN [25]	78	17.6	80	0.035	92.4	0.014
DRIT [26]	164	9.5	132.2	0.016	135.5	0.008
GCGAN [27]	80	8.4	92	0.07	99.8	0.019
CUT [2]	57.3	29.8	79.1	6.9	85.6	2.21
USIS [5]	53.7	44.8	33.2	17.38	27.8	14.06
Ours	**44.1**	**74.7**	**27.3**	**52.1**	**16.8**	**43.8**

scores (mIoU and Acc) across all experimental datasets. In Table 1, a lower FID score and a higher mIoU score indicate better performance. It is evident from the table that our approach achieves significant improvements in both FID and mIoU scores across the three public datasets, demonstrating that our method generates more realistic and semantically well-aligned images than previous unsupervised approaches. Furthermore, the quantitative comparison of our method with recent supervised semantic image synthesis approaches is summarized in Table 2. Most of the baseline models in this table are GAN-based, except for the diffusion-based image synthesis method (SDM [15]) and the Conditional Guided Stable Diffusion Model (ControlNet [1]). The GAN-based models can be further categorized by their improvement directions, including general GANs [6], normalization techniques [11], attention mechanisms [4,28], and discriminator enhancements [7]. Compared to unsupervised approaches, supervised semantic image synthesis methods generally produce higher-quality images by utilizing input data that includes both semantic labels and Gaussian noise. As shown in the table, our method achieves significantly lower FID scores and higher mIoU and Acc scores on the ADE20K and COCO-Stuff datasets compared to previous supervised methods. Although ControlNet [1] leverages a conditional guided stable diffusion model to achieve better generation results on the smaller Cityscapes dataset, its synthesized images are less effective than ours in terms of semantic alignment. Overall, our approach delivers superior quantitative results compared to both unsupervised and supervised methods.

To further highlight the advantages of our proposed method in supervised adversarial learning, we quantitatively compare our approach with other methods using various evaluation metrics. Specifically, we use Learned Perceptual Image Patch Similarity (LPIPS) [30], Peak Signal-to-Noise Ratio (PSNR) [31], and Topological Image Quality (TOPIQ) [32]. LPIPS measures perceptual similarity between images, focusing on the high-level features that align with human visual perception, where lower values indicate closer perceptual similarity. PSNR assesses the structural integrity of images by comparing pixel-level differences, with higher values indicating better reconstruction quality. TOPIQ evaluates overall image quality, including aspects of both visual perception and topological accuracy, where higher scores reflect superior diversity and structural consistency. As shown in Table 3, our approach demonstrates notable advantages in these areas, indicating superior performance in image synthesis. Furthermore, we evaluate the transferability of our method across datasets. Specifically, we use the pretrained model on ADE20K and perform inference on COCO-Stuff, comparing the Image Quality Assessment (IQA) [29] metrics across different methods. IQA provides an overall assessment of image quality by evaluating both structural and perceptual aspects. The results, presented in Table 4, showcase our method's superior performance and generalization ability across different datasets. This further emphasizes the robustness and effectiveness of our approach in diverse scenarios.

These improvements are not just theoretical but have practical implications in real-world tasks. For instance, when our method is applied as a data augmentation tool in semantic segmentation, the enhanced image quality directly translates into better model performance. This is substantiated by the new Table 5, which compares our method with the latest baseline methods for data augmentation in semantic segmentation. In this table, "Ground Truth" represents the baseline performance of the segmentation network trained on existing datasets without augmentation. When used to augment the training data, our method significantly improves performance metrics compared to the baselines. For example, on the Cityscapes dataset, our method reduces the Top-1 error rate, a metric that measures the percentage of the most likely predictions being incorrect, from 19.3% to 18.1%. This indicates a reduction in misclassification. Similarly, the mean Intersection over Union (mIoU), which measures the overlap between predicted and ground truth regions across all classes, increases from 47.2% to 50.4%. Higher mIoU values reflect better segmentation accuracy and alignment with ground truth. Additionally, the pixel Accuracy (Acc), which calculates the percentage of correctly classified pixels, improves from 38.2% to 41.6%, indicating a more accurate pixel-level prediction. These results provide concrete evidence that the gains in image quality and diversity achieved by our method have direct and significant benefits in practical applications, reinforcing its effectiveness as a data augmentation tool.

4.2.1. Human perceptual evaluation

To further validate our method's performance in semantic image synthesis, we conducted a human perceptual evaluation [33] across the Cityscapes, ADE20K, and COCO-Stuff datasets, comparing our approach with the latest baseline methods [1,4]. We employed a rigorous random sampling approach to ensure representativeness and minimize bias: 200 semantic labels were randomly selected from each dataset's validation set using Python's random library, ensuring a diverse and unbiased selection. For participant selection, we randomly chose 20 individuals with expertise in image processing and computer vision, ensuring a broad range of opinions and enhancing the evaluation's reliability. The evaluation was conducted under standardized conditions, with images presented in a randomized order to prevent bias. Statistical analysis was performed to calculate the average probability of our method's images being selected as more photo-realistic, with p-values indicating the statistical significance of the observed differences, as summarized in Table 6. Smaller p-values indicate the robustness and reliability of our method's superior performance compared to the baseline methods.

4.2.2. Result with dual-function discriminator

We compared the discriminator of our method with those of GAN-based benchmarks in terms of network structure, parameters, innovation, and functionality, as shown in Table 7. Previous GAN-based models for semantic image synthesis typically use discriminators like "MultiscalePatchGAN", "Semantics-embedding", "Segmentation-based", and "MultiscaleSESAME". The "MultiscalePatchGAN" from Pix2PixHD [17] is based on PatchGAN [16], while CC-FPSE [18] introduced the "Semantics-embedding" discriminator, which reduces the number of parameters. As shown in Fig. 6, although OASIS employs a segmentation-based discriminator, its main focus is on differentiating real from fake images, which has only an indirect and limited impact on semantic alignment. In contrast, "MultiscaleSESAME" in SAFM [7], built on SESAME [34], incorporates a complex SegNet structure to assist in prediction. Our proposed dual-function discriminator, however, not only distinguishes between real and fake images but also performs multi-class segmentation. It uses fewer parameters than both OASIS and SAFM, and the improvements in semantic alignment are evident in the per-class IoU scores for Cityscapes, as shown in Table 8. Our method achieves higher IoU scores across all object classes, with particularly significant gains for larger classes such as utility poles, traffic lights, buses, trains, and motorcycles, resulting in a +6 obj-mIoU improvement over the state-of-the-art OCGAN [4] model.

Table 2
Quantitative comparison of our method with supervised baseline methods based on image generation scores (FID) and semantic segmentation metrics (mIoU and Acc) across all datasets. 'n/a' indicates that the visual result is not provided on the model's official website. Boldface highlights the best performance.

Method	Cityscapes			ADE20K			COCO-Stuff		
	FID↓	mIoU↑	Acc↑	FID↓	mIoU↑	Acc↑	FID↓	mIoU↑	Acc↑
DAGAN [6]	60.3	66.1	82.6	31.9	40.5	81.6	n/a	n/a	n/a
LGGAN [28]	57.7	68.4	83.0	31.6	41.6	81.8	n/a	n/a	n/a
SPADE [11]	71.8	62.3	81.9	33.9	38.5	79.9	22.6	37.4	67.9
SAFM [7]	49.5	70.4	83.1	32.8	50.1	86.6	24.6	43.3	73.4
SDM [15]	42.1	72.8	93.7	27.5	43.5	82.4	23.5	39.8	68.7
ControlNet [1]	**42.0**	73.1	93.9	27.8	45.6	84.0	21.3	42.7	70.2
OCGAN [4]	43.8	73.8	94.8	28.3	51.2	86.8	17.0	43.2	73.8
Ours	44.1	**74.7**	**95.4**	**27.3**	**52.1**	**87.3**	**16.8**	**43.8**	**74.0**

Table 3
Quantitative comparison of our method with supervised baseline approaches using other evaluation metrics for assessing image quality and diversity.

Method	Cityscapes			ADE20K			COCO-Stuff		
	LPIPS	PSNR	TOPIQ	LPIPS	PSNR	TOPIQ	LPIPS	PSNR	TOPIQ
DAGAN	0.3927	16.191	0.3840	0.5175	11.766	0.5405	n/a	n/a	n/a
LGGAN	0.3830	11.766	0.3526	0.5138	11.840	0.5317	n/a	n/a	n/a
SPADE	0.3953	15.848	0.4103	0.5178	11.731	0.5572	0.6520	11.200	0.5904
SAFM	0.3824	14.095	0.4571	0.5018	12.046	0.5100	0.6346	10.088	0.6103
SDM	0.3812	15.370	0.4594	0.5032	11.921	0.6419	0.5571	10.352	0.6845
ControlNet	0.3857	15.936	0.4826	**0.4963**	12.082	0.6107	0.5492	10.310	0.7061
OCGAN	0.3792	**16.243**	0.4380	0.5031	12.109	0.6741	0.5381	11.063	**0.7119**
Ours	**0.3775**	15.578	**0.5273**	0.5192	**12.173**	**0.6975**	**0.5364**	**11.209**	0.7093

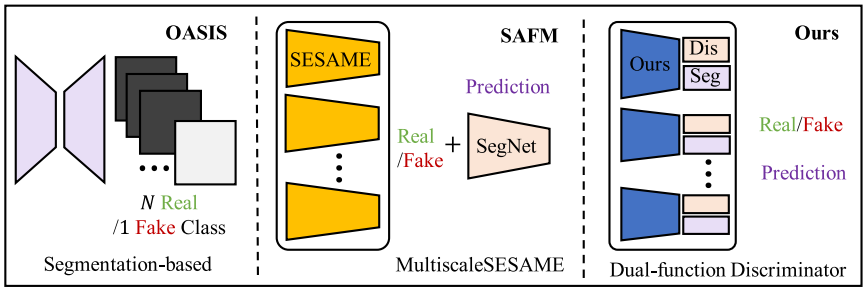


Fig. 6. Comparison of semantic alignment between our method and baseline approaches, highlighting improvements in spatial consistency and structure preservation.

Table 4
The comparison results of cross-dataset transferability, with the metric being Image Quality Assessment (IQA) values.

Metric	SPADE [11]	SAFM [7]	SDM [15]	ControlNet [1]	OCGAN [4]	Ours
IQA [29]	0.75	0.78	0.80	0.82	0.79	**0.85**

Table 5
Comparison of the proposed method with the latest baseline methods as data augmentation tools for semantic segmentation tasks.

Method	Cityscapes		ADE20K		COCO-Stuff	
	Top-1	Top-5	mIoU	Acc	mIoU	Acc
Ground Truth	20.6	5.5	42.0	80.7	34.8	65.8
ControlNet [1]	20.0	5.1	44.1	83.5	36.0	66.9
OCGAN [4]	19.3	4.4	47.2	84.9	38.2	70.1
Ours	18.1	3.9	50.4	86.8	41.6	73.5

Table 6
Human perceptual evaluation. These values reflect the average probability of our method being approved by the workers compared to the baseline methods in image synthesis. p-values indicate the statistical significance of the observed differences.

Method	Cityscapes	ADE20K	COCO-Stuff	p-values
Ours > ControlNet	63.28%	57.36%	57.04%	0.018
Ours > OCGAN	58.17%	51.90%	52.67%	0.026

Table 7
Comparative results of our discriminator with that of the GAN-based benchmarks in terms of structure, number of parameters, innovation and functionality.

Method	Discriminator structure	Parameters	Innovation	True/False	Prediction
LGGAN	MultiscalePatchGAN	5.6	×	✓	×
SPADE	MultiscalePatchGAN	5.6	×	✓	×
OCGAN	MultiscalePatchGAN	5.6	×	✓	×
CC-FPSE	Semantics-embedding	**5.2**	✓	✓	×
OASIS	Segmentation-based	22.3	✓	✓	✓
SAFM	MultiscaleSESAME	33.4	✓	✓	✓
Ours	Dual-function	8.1	✓	✓	✓

4.2.3. Analysis of method efficiency

In Table 9, we present a comparison of the network parameters and model efficiency between our approach and several representative competing models. Our GAN-based model stands out with the smallest total number of trainable parameters, totaling just 92.7 million, thanks to the lightweight design of both the generator and discriminator. This allows us to reduce the number of parameters without compromising the quality of image synthesis. Moreover, we assess the model's performance in terms of inference time and frames per second (fps) on a single GeForce RTX 3090 GPU. The results reveal that our model reduces inference time by more than 3.3% and increases fps by over 3.3% compared to the latest OCGAN [4]. These findings highlight that our method not only achieves superior image synthesis quality but

Table 8
Per-class IoU for the Cityscapes dataset. Classes corresponding to objects, i.e. the object, human, and vehicle groups in Cityscapes, are underlined. The obj-mIoU refers to the mIoU calculated only for these object classes.

Method	road	swalk	build	wall	fence	pole	tlight	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	obj-mIoU
DAGAN	97	80	89	60	53	41	39	46	88	65	92	66	45	89	71	75	57	25	60	56
LGGAN	97	83	89	59	56	42	42	50	89	70	92	69	48	90	72	80	52	38	64	59
SPADE	97	80	88	54	50	40	39	41	88	69	92	66	41	89	64	73	42	29	61	53
SAFM	99	92	**95**	79	80	56	73	65	**97**	91	96	87	68	**98**	81	88	40	56	84	72
SDM	97	82	88	58	62	44	42	50	85	69	85	60	49	89	66	71	47	41	63	57
ControlNet	98	80	87	72	65	42	52	51	87	**91**	79	66	55	89	67	81	**65**	55	65	69
OCGAN	97	90	92	**79**	78	46	60	58	93	89	89	80	58	93	84	73	47	60	72	73
Ours	**99**	**92**	95	79	**82**	**58**	**77**	**73**	97	90	**96**	**87**	**72**	97	**88**	**93**	57	**66**	**89**	**79**

Fig. 7. High-resolution image synthetic results on the Cityscapes dataset. Our approach generates more photo-realistic and semantically well-aligned images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 9
Comparison of network parameters and model efficiency among different methods. The best performance in each category is highlighted in **bold**.

Method	Network parameters (M)			Model efficiency	
	Total	Generator	Discriminator	Inference time (s) ↓	FPS ↑
LGGAN [28]	117.2	111.6	5.6	0.2650	3.77
SPADE [11]	98.6	93.0	5.6	0.6062	1.63
CC-FPSE [18]	133.3	128.1	**5.2**	0.0913	10.95
OASIS [19]	93.4	**71.1**	22.3	0.0875	11.43
SAFM [7]	129.6	96.2	33.4	0.0963	10.38
OCGAN [4]	97.5	91.9	5.6	0.0879	11.38
Ours	**92.7**	84.3	8.1	**0.0850**	**11.76**

Table 10
Similarity matching results of the average power spectrograms between Ground Truth (GT) images and those synthesized by deep learning methods.

Power spectrograms	ORB similarity	Hist similarity
GT vs. SPADE [11]	0.64330	0.59132
GT vs. SAFM [7]	0.71308	0.62461
GT vs. ControlNet [1]	0.78143	0.71036
GT vs. OCGAN [4]	0.81325	0.78934
GT vs. Ours	**0.87059**	**0.84062**

also demonstrates enhanced computational efficiency, outperforming state-of-the-art models in both aspects.

4.3. Qualitative results

Figs. 7, 8, and 9 present the qualitative comparison between our proposed method and the competing approaches on the Cityscapes, ADE20K, and COCO-Stuff datasets. Our findings show that the images synthesized by our method not only exhibit superior perceptual quality but also more closely match the ground truth images in terms of overall color and texture distribution. Notably, the complex real-world scenes generated by our method demonstrate significant improvements on the Cityscapes dataset. Although OCGAN [4] is the current state-of-the-art GAN-based approach, its synthesized images tend to be overly bright and even exhibit color distortion. In contrast, our method produces photo-realistic images that faithfully adhere to the input semantic labels, and it excels in generating challenging scenes with high image fidelity.

4.3.1. Average power spectrogram

It reflects the distribution of energy from low to high frequencies in images, and the extent to which low-frequency energy is concentrated in the center of the spectrum. Therefore, we furthermore conducted a qualitative comparison of the average power spectrum for synthesized images on the ADE20K dataset, whose result is shown in Fig. 10. The average power spectrum of the images generated by the competing methods presents more significant peaks and distortions. Moreover, the visual comparison in terms of color, shape and texture also confirms that the average power spectrum is more similar between the ground truth images and the images generated by our approach. Table 10 reports the similarity matching result of the average power spectrograms between ground truth images and synthesized images. Specifically, the similarity of the two average power spectrograms is calculated by the ORB [35] and histogram algorithms. And the higher the calculated value, the more similar. The similarity matching results also helpful to confirm that the distribution of the images synthesized by our method is closer to the ground truth images.

4.3.2. Visualization of predicted semantic maps

Following OASIS [19], we employ a pre-trained segmentation network to produce predicted semantic maps from the synthesized images. Details of the specific pre-trained networks used for each dataset are provided in Section 4.1, such as DRN-D-105 [36] for Cityscapes, UperNet101 [37] for ADE20K. Fig. 11 shows a visual comparison between the semantic labels and the predicted semantic maps generated by our method versus those produced by the latest OCGAN [4]. Our results indicate that the images generated by our method are of higher quality, and the predicted semantic maps are more accurately aligned with the

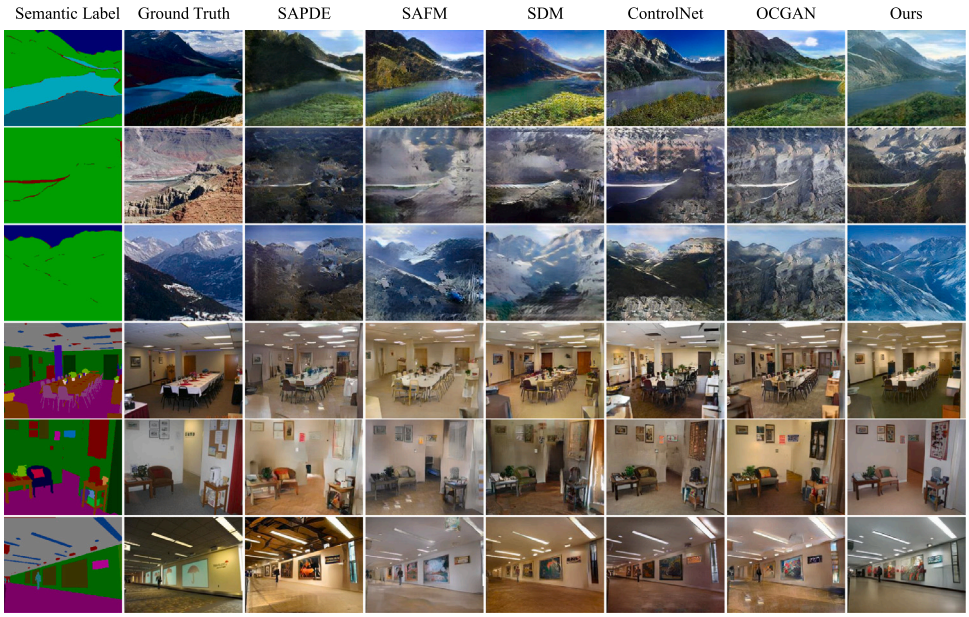


Fig. 8. Visual comparison results on the ADE20K dataset. Though very diverse categories and small structures, our method can work well and generate high-fidelity results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

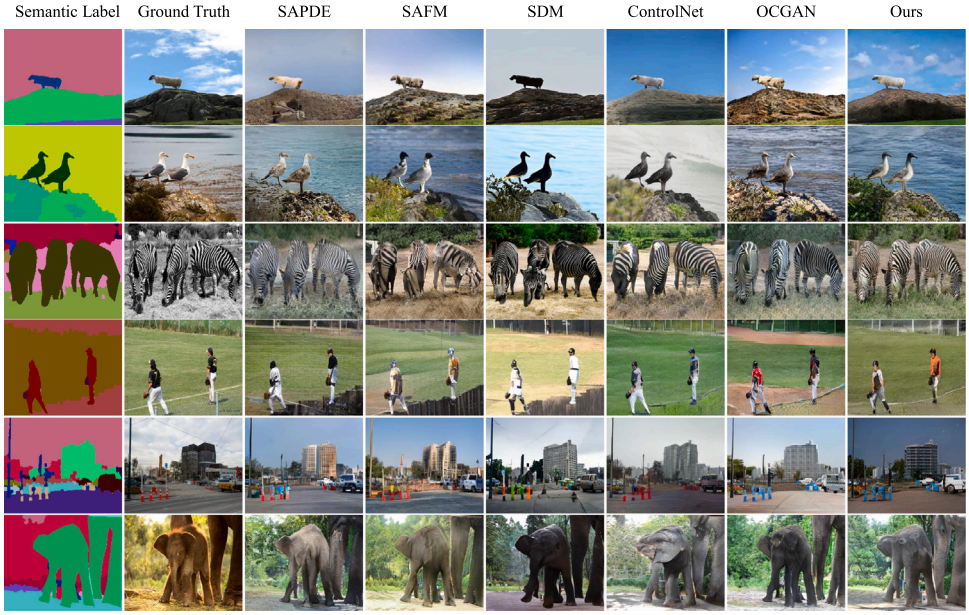


Fig. 9. Visual comparison results on the challenging COCO-Stuff dataset. It shows that images synthesized by our approach are more realistic than baselines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

given semantic labels compared to those from OCGAN. Particularly, our method delivers improved predictions for small objects, such as "pole" and "trolley" in the Cityscapes dataset. This highlights the effectiveness of incorporating a segmentation function into our discriminator to enhance semantic alignment.

4.3.3. Visualization of learned attention heatmaps

To analyze the proposed CSAM (Fig. 4), we randomly selected its position in the generator (e.g., at DL_1) and displayed the input and output feature heatmaps in columns 1 and 6 of Fig. 12. The output heatmap from CSAM not only reduces channel dimensions (from 96 to 64) but also enhances attention intensity and object contours. For example, the grass, bed, and chair, highlighted by black dashed lines in column 6, show brighter heatmaps and clearer contours compared to column 1. This suggests that CSAM improves feature quality for synthesizing high-quality images. As shown in Fig. 4, CSAM combines channel-wise and spatial-wise attention. The output heatmap is a fusion of these attentions. For the channel attention heatmap (column 4), we display attention maps from two channels (N_1th and N_2th) in columns 2 and 3. Differences in channel attention weights are evident; for instance, the 63th channel map is brighter than both the 6th, indicating it contains more significant information. Column 5 displays the spatial attention heatmap, highlighting the importance of different regions by visualizing the attention weights across them. Moreover, the output heatmaps (in column 6) are more focused on significant information than those from the channel (column 4) and spatial attention maps (column 5).

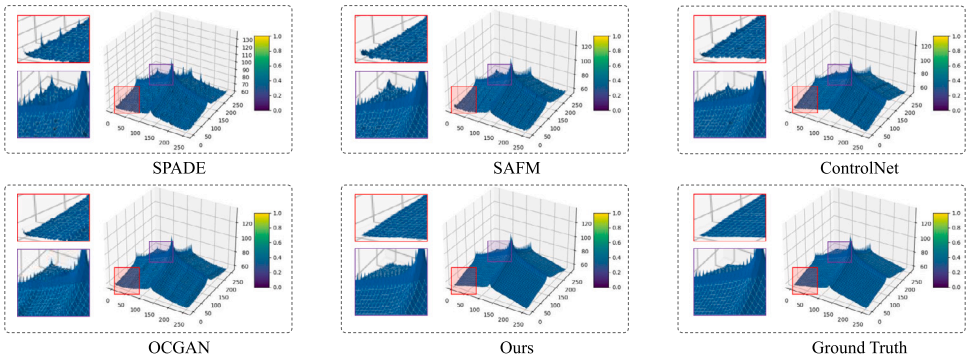


Fig. 10. Qualitative comparison of the average power spectrograms drawn from ground truth images and synthetic images on the ADE20K dataset. Coordinate axes are on a logarithmic scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

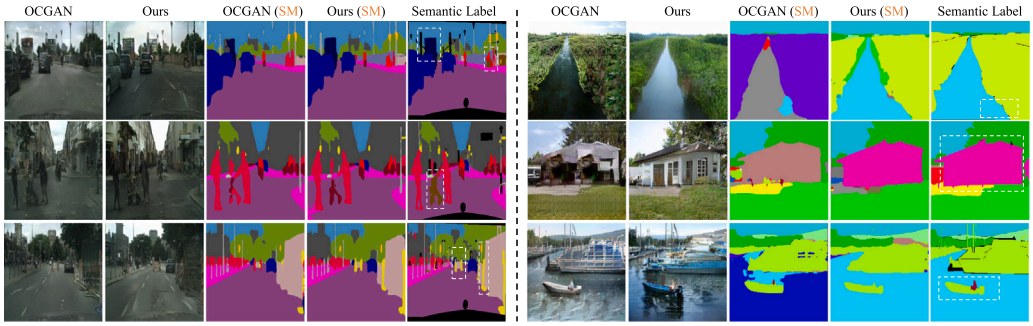


Fig. 11. Visualization of predicted semantic maps for the synthesized images compared with those from SAFM on the Cityscapes (left) and ADE20K (right) datasets. Semantic image synthesis results of our method and latest SAFM. "SM" is an abbreviation for Semantic Map.

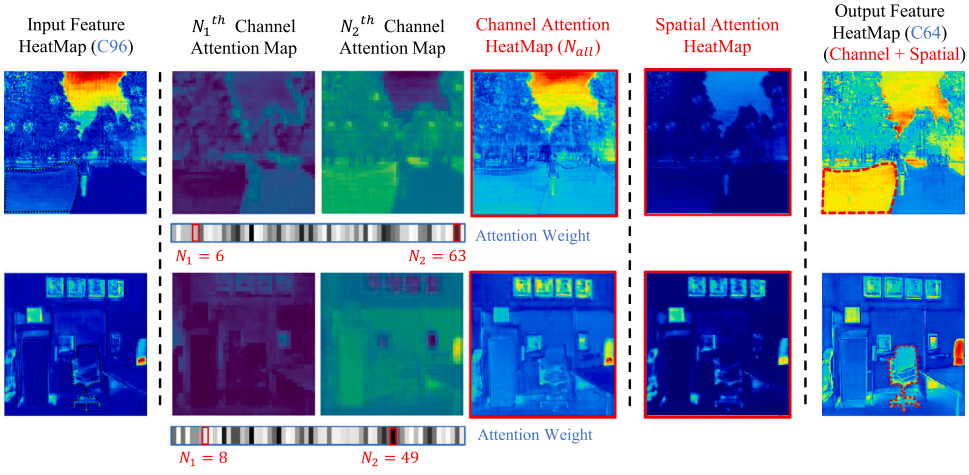


Fig. 12. Heatmap visualization of the Channel and Spatial Attention Mechanism (CSAM) on ADE20K. Columns 1 and 6 show the feature heatmaps of the input and output of CSAM (as illustrated in Fig. 4). The output heatmap (column 6) results from the fusion of the channel attention heatmap (column 4) and the spatial attention heatmap (column 5). The channel attention heatmap consists of multiple channels, with the attention weight for the N_1th channel represented in a color scale, where darker colors indicate stronger attention. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4. Ablation study

We propose VD-GAN, a novel GAN-based approach that enhances both image generation and semantic segmentation. An ablation study will be conducted to validate the effectiveness of the novel module in semantic image synthesis.

4.4.1. Ablation study on the generator network

Our generator network is based on a variational U-Net structure, consisting of an encoder and decoder. The encoder uses Swin Transformer blocks and a patch embedding layer, while the decoder incorporates Swin Transformer blocks, a patch expanding layer, and a Conditional Residual Attention Module (CRAM). This architecture combines the Swin Transformer (SwinT) with CNNs, leveraging pre-trained (PT) weights for improved performance. We conducted an ablation study (denoted as "*SwinT w/ PT + CNN*") to demonstrate the effectiveness of our approach, as detailed in Table 11. The experimental setups include: "*RNN-UNet*", a generator that integrates Recurrent Neural Networks (RNN) into the UNet architecture to enhance feature extraction by capturing temporal dependencies; "*CNN-UNet*", a fully convolutional generator with an encoder and decoder structure incorporating CRAM; and "*SwinT w/ PT*", our generator without the CRAM module, which fuses shallow and deep features. The comparison of our method with these alternatives highlights the advantages of

Table 11
Ablation study on the generator network.

Method	LPIPS↓	FID↓	mIoU↑	Acc↑
RNN-UNet	0.5402	47.9	68.5	88.3
CNN-UNet	0.4274	48.4	69.2	90.8
SwinT w/ PT	0.3927	45.1	72.0	94.7
SwinT w/o PT + CNN	0.3953	46.3	71.9	93.2
SwinT w/ PT + CNN	**0.3775**	**44.1**	**74.7**	**95.4**

Table 12
Ablation study on the conditional residual attention module (CRAM) and the conditional batch normalization (CBN). The bold font indicates our settings.

Config.	CRAM	Conv [16]	CBN	BN [38]	SPADE [11]	FID	mIoU	Acc
A1	✓	×	✓	×	×	44.1	**74.7**	**95.4**
A2	✓	×	×	✓	×	45.0	73.1	94.8
A3	✓	×	×	×	✓	**44.0**	74.5	95.0
A4	×	✓	✓	×	×	44.7	73.6	94.5

Table 13
Ablation study of our CSAM on Cityscapes. "CAM" and "SAM" represents the channel-wise and spatial-wise attention operation in the proposed CSAM, respectively. For all metrics except FID, higher is better.

	Settings	FID↓	mIoU↑	Acc↑
B1	Ours	**44.1**	**74.6**	**95.4**
B2	B1-CAM	45.1	73.9	95.3
B3	B1-SAM	44.7	74.2	95.4
B4	B2-SAM	47.0	73.5	94.9

a Transformer-based network for semantic image synthesis and the improved performance from combining the Transformer with CNNs. We also evaluated a version of our network with randomly initialized parameters ("SwinT w/o PT + CNN"), and the results underline the critical role of pre-trained weights in the Swin Transformer.

4.4.2. Ablation study on the CRAM and CBN

In order to compensate for the information loss resulting from the encoding stage, the extracted low-level features are concatenated to the high-level features output by the upper layer of the decoder. Nevertheless, the concatenation operation acquires higher dimensional features with spatial and semantic information. To reduce the feature dimensionality and extract important information, we propose a novel conditional residual attention module, abbreviated as CRAM. In order to highlight the superiority of our proposed CRAM, its baseline variant utilizes a conventional 1×1 convolution block as a comparison. Inspired by BN [38], we also design a Conditional Batch Normalization (CBN) layer in the structure of CRAM. To evaluate the importance of the innovated CBN, we employ two normalization layers available for replacement: a common Batch Normalization [38] layer (also known as "BN") and an advanced SPatially-Adaptive (DE)normalization [11] (denoted as "SPADE"). As shown in Table 12, the selection of CRAM and SPADE only performs well in the image generation score (FID), while both CRAM and CBN used in our generator network obtain outstanding manifestation in the semantic segmentation scores (mIoU and Acc).

4.4.3. Ablation study on the CSAM

To evaluate the effectiveness of the proposed Channel and Spatial Attention Mechanism (CSAM) in VD-GAN, we conduct an ablation study by comparing the full VD-GAN model with three variants, each of which progressively removes specific components from the CSAM structure (as shown in Fig. 4). The variants are as follows: (B1) "Baseline", which represents the VD-GAN model with the complete CSAM structure; (B2) "B1-CAM", where the channel-wise attention operation is removed; (B3) "B1-SAM", where the spatial-wise attention operation is excluded; and (B4) "B2-SAM", which omits the entire CSAM structure, thus lacking both channel and spatial attention mechanisms. The ablation study results, presented in Table 13, demonstrate the crucial role of both the channel-wise and spatial-wise attention operations in enhancing the quality of image synthesis conditioned on semantic labels.

Table 14
Comparison of the discriminator's performance on the Cityscapes dataset with different $Scale_D$ settings, including a multi-scale discriminator.

Method	FID	mIoU	Acc
$Scale_D = 1$	45.6	73.4	95.1
$Scale_D = 2$	44.1	74.6	95.4
$Scale_D = 3$	46.1	73.2	95.2
Multi-scale	45.5	69.4	94.8

4.4.4. Ablation on input information utilization

To assess the effectiveness of our approach in utilizing input information compared to previous GAN models, we compared feature heatmaps at critical network stages—specifically, the encoder layer 4 (EL_4) and decoder layer 2 (DL_2). As shown in Fig. 13, the green boxes represent heatmaps from earlier GAN models, where input data is directly fed into the UNet structure, while the red boxes show heatmaps from our method, which uses semantic enhancement through one-hot encoding. This comparison highlights that our method achieves clearer and more distinct feature activations at these stages, demonstrating a more effective utilization of input information. Our approach enhances semantic representation and compensates for potential information loss during encoding, resulting in improved network performance and a better extraction and use of input data.

4.4.5. Ablation on the discriminator network

In addition, our discriminator adopts a multi-scale strategy to improve the training of the network. To verify the effectiveness of this approach, we conducted experiments with discriminators set at different scales, while keeping other parameters constant. The FID, mIoU, and Acc metrics on the Cityscapes dataset (Table 14) indicate that a two-scale discriminator, which operates at both the original scale and a downsampled scale of 1/2, leads to higher quality results and reduced training times. Additionally, we compared the performance of our dual-function discriminator against a conventional multi-scale discriminator [17]. The results show that our dual-function discriminator significantly improves the photo-realism of the synthesized images and enhances semantic alignment, as evidenced by the metric scores.

5. Conclusion

In this paper, we propose a novel VD-GAN method to improve image quality and semantic consistency in image synthesis. By leveraging a U-Net-based generator with attention modules, the dual-function discriminator ensures consistency between the generated images and semantic labels. Experimental results demonstrate state-of-the-art performance across three standard datasets, and Fig. 14 showcases our approach's generalizability on the Helen [39] dataset, along with its potential in artistic applications such as multi-style image synthesis and semantic control synthesis. However, we recognize that the Helen dataset, consisting of 2330 facial images with detailed and consistent annotations, may introduce biases related to facial features, age, or expressions, which could impact model performance, particularly when applied to more diverse or complex scenarios. Additionally, ethical considerations regarding the potential misuse of generated images or biases in synthetic content must be carefully addressed to ensure responsible application of our approach. Our method also faces challenges in handling complex scenes with intricate backgrounds and fine details. For instance, scenarios involving ambiguous or incomplete semantic information, low-quality input images, or complex background details may lead to less accurate or realistic image generation. These failure cases

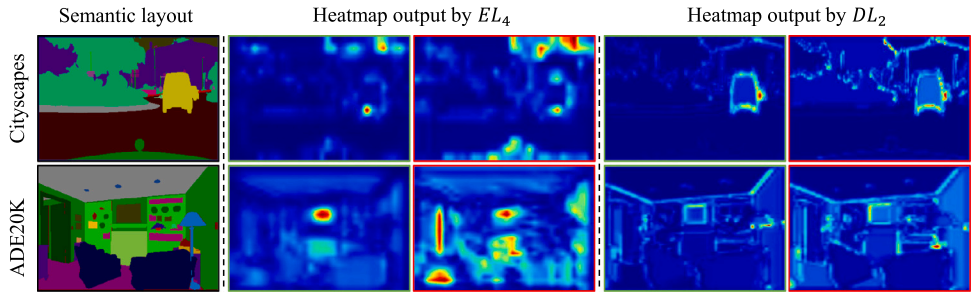


Fig. 13. Comparison of feature heatmaps at different stages of the generator network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

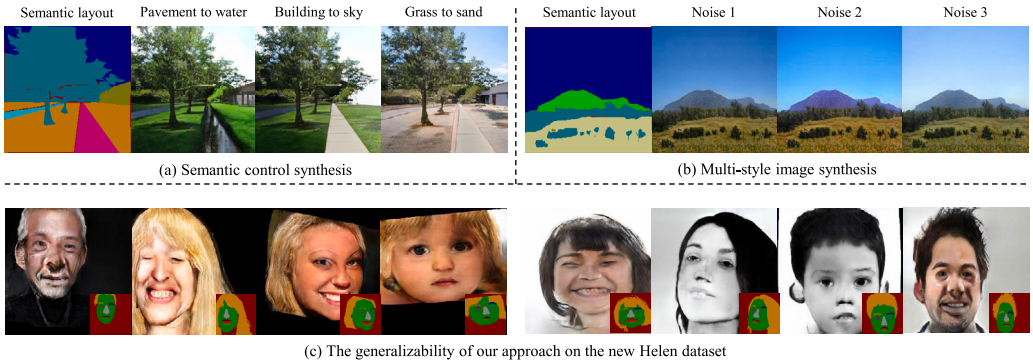


Fig. 14. (a–b)Artistic applications of multi-style image synthesis and semantic control synthesis based on our method.(c) The generalizability of our approach across different domains.

highlight the need for further refinement in handling detailed features and more intricate scenes. To address these challenges, future work will focus on mitigating dataset biases, enhancing the model's robustness, and improving its ability to handle such complexities. Moreover, we aim to explore real-world deployment opportunities in industries such as entertainment, virtual reality, and digital media, ensuring that our approach meets the specific needs of these applications while adhering to high ethical standards.

CRediT authorship contribution statement

Aihua Ke: Writing – original draft, Methodology. **Bo Cai:** Data curation, Conceptualization. **Yujie Huang:** Visualization, Software. **Jian Luo:** Data curation, Conceptualization. **Yaoxiang Yu:** Visualization, Resources. **Le Li:** Data curation, Project administration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Bo Cai reports financial support was provided by National Natural Science Foundation of China.

Acknowledgments

The work described in this paper was funded by National Natural Science Foundation of China: "Research on Key Technologies of Highly Available Indoor Positioning for Wide Area Signal of Opportunity Environment" (Fund No. 61971316).

Data availability

Data will be made available on request.

References

[1] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3836–3847.
[2] T. Park, A.A. Efros, R. Zhang, J.-Y. Zhu, Contrastive learning for unpaired image-to-image translation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, Springer, 2020, pp. 319–345.
[3] A. Ke, Y. Huang, J. Yang, B. Cai, Text-guided image-to-sketch diffusion models, Knowl.-Based Syst. 304 (2024) 112441.
[4] G. Liu, A. Ke, X. Wu, H. Zhang, GAN with opposition-based blocks and channel self-attention mechanism for image synthesis, Expert Syst. Appl. 246 (2024) 123242.
[5] G. Eskandar, M. Abdelsamad, K. Armanious, B. Yang, USIS: Unsupervised semantic image synthesis, Comput. Graph. (2023).
[6] H. Tang, S. Bai, N. Sebe, Dual attention gans for semantic image synthesis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1994–2002.
[7] Z. Lv, X. Li, Z. Niu, B. Cao, W. Zuo, Semantic-shape adaptive feature modulation for semantic image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11214–11223.
[8] J. Janosovits, Cityscapes tl++: Semantic traffic light annotations for the cityscapes dataset, in: 2022 International Conference on Robotics and Automation, ICRA, IEEE, 2022, pp. 2569–2575.
[9] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ade20k dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 633–641.
[10] C.-H. Lee, Z. Liu, L. Wu, P. Luo, Maskgan: Towards diverse and interactive facial image manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5549–5558.
[11] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.
[12] T. Fontanini, C. Ferrari, G. Lisanti, M. Bertozzi, A. Prati, Semantic image synthesis via class-adaptive cross-attention, IEEE Access (2025).
[13] X. Qi, Q. Chen, J. Jia, V. Koltun, Semi-parametric image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8808–8816.
[14] M. Xu, J. Lee, S. Yoon, H. Kim, D.S. Park, Variation-aware semantic image synthesis, Image Vis. Comput. 142 (2024) 104914.
[15] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, H. Li, Semantic image synthesis via diffusion models, 2022, arXiv preprint 2207.00050.

[16] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[17] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8798–8807.
[18] X. Liu, G. Yin, J. Shao, X. Wang, et al., Learning to predict layout-to-image conditional convolutions for semantic image synthesis, Adv. Neural Inf. Process. Syst. 32 (2019).
[19] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, A. Khoreva, OASIS: only adversarial supervision for semantic image synthesis, Int. J. Comput. Vis. 130 (12) (2022) 2903–2923.
[20] J. Yang, A. Ke, Y. Yu, B. Cai, Scene sketch semantic segmentation with hierarchical transformer, Knowl.-Based Syst. 280 (2023) 110962.
[21] N.A. Namin, E. Garaaghaji, M. Rezaei, M.Z. Lighvan, Light weight semantic segmentation: A modified DDRNET approach trained on cityscapes and COCO-stuff datasets for efficient image analysis, in: 2023 7th International Symposium on Innovative Approaches in Smart Technologies, ISAS, IEEE, 2023, pp. 1–5.
[22] D.A. Chan, S.P. Sithungu, Evaluating the suitability of inception score and fréchet inception distance as metrics for quality and diversity in image generation, in: Proceedings of the 2024 7th International Conference on Computational Intelligence and Intelligent Systems, 2024, pp. 79–85.
[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[24] Y. Wang, Z. Xiao, G. Cao, A convolutional neural network method based on adam optimizer with power-exponential learning rate for bearing fault diagnosis, J. Vibroeng. 24 (4) (2022) 666–678.
[25] S. Benaim, L. Wolf, One-sided unsupervised domain mapping, Adv. Neural Inf. Process. Syst. 30 (2017).
[26] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, M.-H. Yang, Diverse image-to-image translation via disentangled representations, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 35–51.
[27] H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang, D. Tao, Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2427–2436.
[28] H. Tang, L. Shao, P.H. Torr, N. Sebe, Local and global GANs with semantic-aware upsampling for image generation, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) (2022).
[29] K. Ding, K. Ma, S. Wang, E.P. Simoncelli, Image quality assessment: Unifying structure and texture similarity, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2020) 2567–2581.
[30] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[31] J. Korhonen, J. You, Peak signal-to-noise ratio revisited: Is simple beautiful? in: 2012 Fourth International Workshop on Quality of Multimedia Experience, IEEE, 2012, pp. 37–38.
[32] C. Chen, J. Mo, J. Hou, H. Wu, L. Liao, W. Sun, Q. Yan, W. Lin, Topiq: A top-down approach from semantics to distortions for image quality assessment, IEEE Trans. Image Process. (2024).
[33] Y. Wang, L. Qi, Y.-C. Chen, X. Zhang, J. Jia, Image synthesis via semantic composition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13749–13758.
[34] E. Ntavelis, A. Romero, I. Kastanis, L. Van Gool, R. Timofte, Sesame: Semantic editing of scenes by adding, manipulating or erasing objects, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, Springer, 2020, pp. 394–411.
[35] Y. Xie, Q. Wang, Y. Chang, X. Zhang, Fast target recognition based on improved ORB feature, Appl. Sci. 12 (2) (2022) 786.
[36] F. Yu, V. Koltun, T. Funkhouser, Dilated residual networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 472–480.
[37] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, Unified perceptual parsing for scene understanding, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 418–434.
[38] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, PMLR, 2015, pp. 448–456.
[39] Q. Zheng, J. Deng, Z. Zhu, Y. Li, S. Zafeiriou, Decoupled multi-task learning with cyclical self-regulation for face parsing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 4156–4165.