Semantic Segmentation Mask-Guided Diffusion Models: A Pathway to Enriched Datasets in Autonomous Systems

Katica Bozsó, András Béres, Bálint Gyires-Tóth
Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics,
1111 Budapest, Műegyetem rkp. 3., Budapest, Hungary,
Email: katica.bozso@gmail.com, {beres,toth.b}@tmit.bme.hu

***Abstract*—In the autonomous vehicle industry, deep learning models are critically dependent on the balance and variety of training data. Achieving this balance is particularly challenging due to the scarcity of data in rare scenarios, such as unique weather conditions or specific traffic configurations. Deep learning-based methods, particularly those within the emerging field of generative artificial intelligence (AI), hold potential for advanced solutions. A key development in this domain is the diffusion-based approach, capable of generating images from a random noise distribution. Predominantly, these models utilize a 'text2image' methodology, enabling the generation of images with text prompts. However, despite their advanced capabilities, these models do not yet provide complete explicit control over the generated content, particularly in terms of the relative positioning of objects within images.**

This research explores the use of a semantic segmentation-based control mechanism within a generative diffusion model, focusing on its application to the automotive domain. With the integration of this mechanism, the model facilitates the creation of diverse and contextually relevant self-driving scene setups, thus enriching the datasets used for comprehensive training in autonomous vehicles. In addition to assessing the quality of generation, the impact of these enriched datasets was also evaluated using a semantic segmentation network, which is essential for advanced driver-assistance systems (ADAS). The study compares the network's performance when trained on the original dataset versus an augmented one that includes model-generated images. The evaluation highlights the practical benefits of applying semantic segmentation guidance in this specific domain.

***Index Terms*—diffusion models, guidance, semantic segmentation, generative AI, data enrichment, ADAS**

I. Introduction

Deep neural networks have revolutionized image generation, finding widespread application in fields such as arts [1], entertainment, medical science [2], and the development of autonomous driving systems [3] [4]. A particularly captivating branch of generative AI is the emergent diffusion-based [5] approach. This method hinges on training a model adept at noise prediction, capable of iteratively crafting images from a standard noise distribution during inference. As many advancements have aimed to enhance the controllability of image generation, cutting-edge solutions now mostly employ a 'text2image' approach, enabling systems to generate images guided by textual prompts [6]. However, they face challenges in accessibility due to their complexity and limitations in precise content control, especially in object positioning within images. This limitation is particularly relevant in autonomous driving, where specific and diverse training data is crucial, yet challenging to acquire.

In response to these challenges, this research introduces a semantic segmentation mask-guided diffusion model tailored for enriching datasets in self-driving environments. By leveraging semantic segmentation maps, this model offers pixel-level control over image content, enabling the generation of complex scene setups that text-based descriptions alone cannot adequately capture. This approach not only facilitates the use of existing semantic segmentation masks for data enrichment, but also allows for easy modification of these masks to generate varied scenarios, enhancing the realism and applicability of training datasets. Utilizing the Berkeley Deep Drive dataset[1], this study demonstrates the model's effectiveness in scene control and its potential for scalability and user-friendly application in generative image modeling.

II. Related Work

A. Semantic Segmentation-guided Diffusion Models

In the landscape of semantic segmentation-guided diffusion models, several notable methods have already been developed. SDEdit [7] offers a novel approach by using stochastic differential equations (SDEs) for image synthesis and editing. This framework iteratively adds and then removes noise to balance between user input faithfulness and image realism, without requiring task-specific training or inversions. ControlNet [8] presents another unique neural network structure that controls large diffusion models to support additional input conditions. It uses a 'trainable copy' and a 'locked copy' of the model, connected with a special type of convolution layer, to learn task-specific conditions robustly, even with small datasets.

The research presented in this work has been supported by Continental Automotive Hungary Ltd.

[1]https://bdd-data.berkeley.edu (access date: 2023.12.17.)

This method mainly utilizes large pretrained Stable Diffusion model and extends it with another controller network. A third approach is the Semantic Diffusion Model (SDM) [9], which processes the semantic layout and noisy image separately in a U-Net structure. This method enhances image quality and semantic relevance by injecting the semantic layout into the decoder via multi-layer spatially-adaptive normalization operators and employs a classifier-free guidance strategy during the sampling process. Differing from these, our approach solely extends a U-Net architecture with an additional convolutional input for mask guidance. This integration allows for precise control over object placement and arrangement in generated images, particularly crucial for complex scenes in autonomous driving.

III. Method

In addressing data enrichment for the automotive domain, our proposed method extends a standard U-Net architecture by adding an extra convolutional block to process one-hot encoded semantic segmentation masks as condition to data generation. This enhancement allows for a more detailed scene depiction, achieved by adding mask channels to the noised input image channels in the initial convolutional layers and propagating them through skip-connections to deeper layers. The approach utilizes images and annotations from the Berkeley Deep Drive dataset, converted into one-hot encoded semantic segmentation masks, to facilitate a nuanced and contextually rich training environment for autonomous driving systems.

A. Dataset

The Berkeley Deep Drive dataset is a large-scale, diverse urban dataset. For this research, 8,000 semantic segmentation annotations were processed, and the original images, with a resolution of 1280x720, were resized to 128x128 pixels after center cropping. This preprocessing resulted in 7,000 training images with corresponding masks and 1,000 image-mask pairs reserved for testing. This dataset, known for its diversity, provides a robust foundation for developing and evaluating the proposed extended U-Net model.

Moreover, in the course of preprocessing, masks for the 19 available classes were generated and saved using a colorbook. This colorbook (Figure 1), which establishes a color-coded scheme corresponding to each class, facilitates the later generation of masks for use in model evaluation and inference.

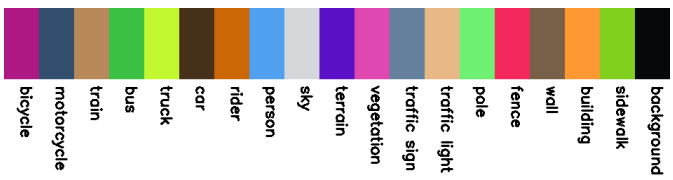

Fig. 1: Class-color mappings

B. Model Design

The design of the proposed model is a tailored U-Net architecture (see Figure 2). Both the input tensor (noisy image) and the one-hot encoded segmentation mask are passed through separate initial convolutional blocks, before adding their outputs channel-wise. This addition enables the utilization of Classifier-Free Guidance [10], even if no guidance is passed, the channel numbers will remain consistent. In the downsampling path, each 'Down' module combines max pooling for spatial reduction and a DoubleConv structure. This DoubleConv setup includes two convolutional layers, interspersed with Group Normalization and GELU activation functions, adding depth and non-linearity to the feature extraction process. The bottleneck further processes the abstracted features through a series of DoubleConv layers. In the upsampling path, 'Up' modules use bilinear upsampling for spatial enlargement, followed by convolutional layers that integrate skip connections from the downsampling path to preserve spatial details. Each 'Up' and 'Down' module incorporates an embedding layer with SiLU activation, enabling the integration of external time information (timestep of the input). The model culminates with a convolutional output layer to map the high-dimensional features to the desired output. To aid more advanced feature extraction and representation learning, two attention blocks were used in the network - both consisting of multi-head attention and linear layers combined with normalization. The potential benefits of augmenting the architecture with additional attention blocks were considered; however, adherence to minimalism was prioritized in this study.

This model diverges from standard designs by embedding the semantic segmentation mask directly into the architecture, though an initial convolutional block. The model was trained on 128x128 images, resulting in a total parameter count of 90,816,131, ensuring a balance between model complexity and computational feasibility.

C. Training

The model was trained exclusively on a single NVIDIA V100 GPU over a period of two and a half days. An AdamW optimizer [11] with a learning rate of 0.0003 was employed. While the training loss plateaued after approximately 300 epochs, indicating minimal gains in traditional loss reduction, the visual fidelity of the results continued to improve, peaking around 550 epochs. This suggests that the model benefits from extended training periods for subtle optimization beyond what is captured by the loss metric alone.

D. Inference

In our study, we developed a pipeline specifically designed for custom image generation. A key aspect of this pipeline is the requirement for users to supply a designated input folder containing semantic segmentation masks. Users must also specify the desired number of images to be generated from each input. The dataloading and subsequent processes are efficiently managed by the pipeline, which is capable of generating unseen images guided by the provided masks. This

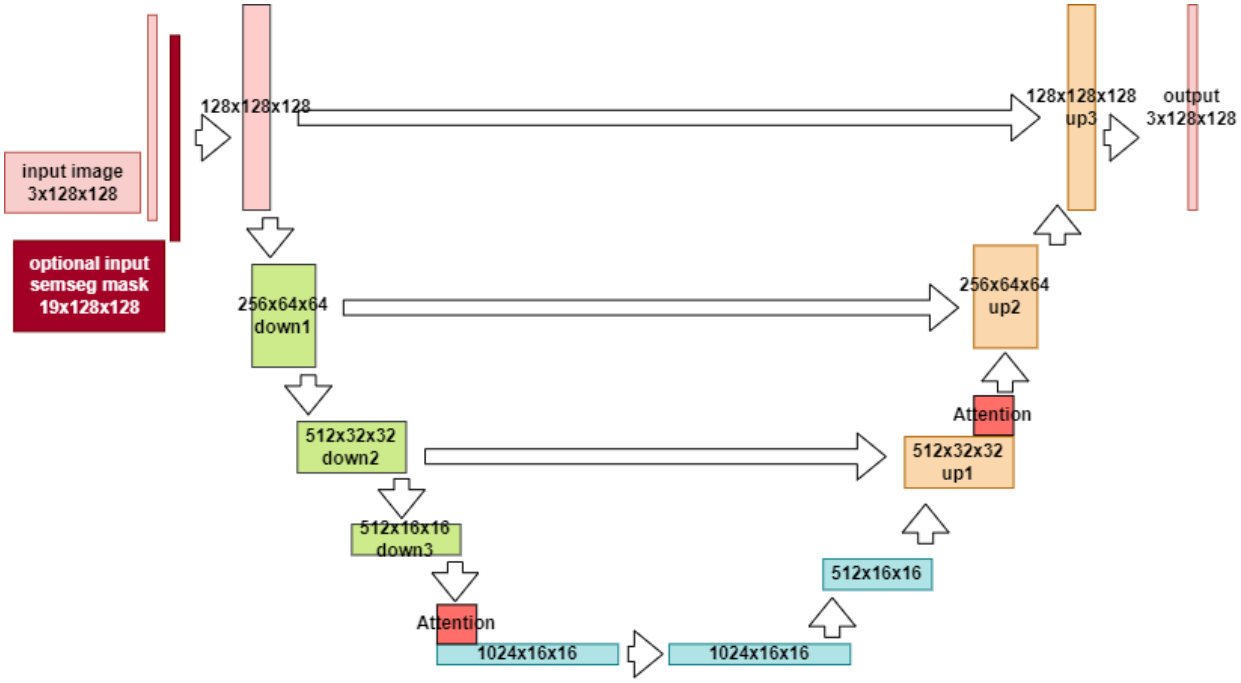

Fig. 2: The utilized U-Net-like architecture

approach ensures that the pipeline is both flexible and user-oriented, enabling the creation of images that precisely meet the specific needs and criteria established by the user.

IV. Experiments

This section delves into the evaluation of our method, conducted through a multi-faceted approach. Initially, a quantitative assessment was undertaken to investigate the generative quality of the results and to gauge their impact on advanced driver-assistance systems. Complementing this, visual experiments were designed to encompass a variety of use-cases, offering a broader perspective on the method's applicability and effectiveness. This dual approach, combining quantitative and qualitative analyses, provides a holistic understanding of the method's performance and its potential implications in real-world scenarios.

A. Quantitative Evaluation - Generation Quality

Using a set of 1000 test images, an equivalent number of novel images were synthesized. To evaluate the quality of these generated images, a pretrained Mask2Former [12] semantic segmentation network was employed. Semantic segmentations were predicted for both the original test images and their corresponding synthesized counterparts. This allowed for the comparison of segmentation consistency, resulting in a notable Intersection over Union (IoU) average of 35.7% and a pixel accuracy of 81.7%. Furthermore, quantitative assessments were conducted using Structural Similarity Index Measure (SSIM), Fréchet Inception Distance (FID) and Kernel Inception Distance (KID), yielding the following results:

TABLE I: Quantitative metrics comparing original and generated images

SSIM	FID	KID	IoU-avg	IoU-PixelAcc
0.6588	1.0636	0.0105	35.7%	81.7%

B. Quantitative Evaluation - Impact on ADAS

To evaluate the impact of our generative method on ADAS, we trained a DeepLabv3 [13] semantic segmentation network on the original dataset of 7000 images (training split), followed by retraining with an augmented set that included 1200 additional synthesized images. These new images were created by sampling 1200 masks from the original dataset, thereby enriching the diversity and complexity of the training data. The key focus of this evaluation was to quantify improvements in the semantic segmentation network's performance. Post augmentation, the validation mean IoU of the network improved from 40.0% to 40.3%, demonstrating the effectiveness of using synthesized imagery in enhancing the robustness and adaptability of ADAS systems. We believe that there is an avenue for larger improvements in the future, by generating an order of magnitude more synthetic images for augmenting the training dataset.

C. Qualitative Evaluation

As described in the previous subsection IV-A, masks were sampled from the test set and novel images were generated using these masks as guidance. (See Figure 3, 4, and for further examples, see Appendix 10). The visual results were in alignment with the structural prerequisites defined by the masks, yielding faithful traffic data samples.

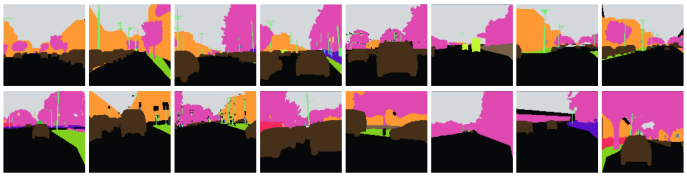

Fig. 3: Test masks

Fig. 4: Generated images for corresponding input masks

Although the test set featured a variety of scene setups, the aspect of custom controllability within our framework is better showcased by actively modifying these masks. To illustrate this, we present an example consisting of an original image, its corresponding initial mask, and the generated image before any alteration on Figure 5.

Fig. 5: Original image, mask and generated image

After modifying the mask by adding an extra car with the corresponding brown color, the structure of the generation changed as expected, with the object inserted into the appropriate place (Figure 6).

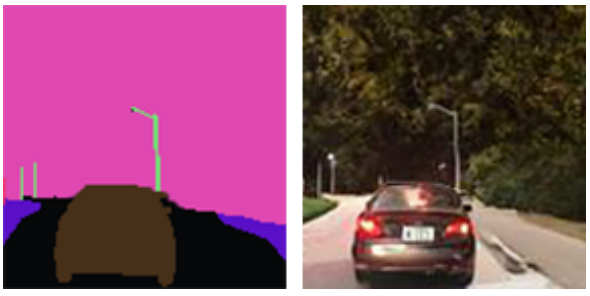

Fig. 6: Modified mask and generated image

The most intriguing approach may yet be the third option. Users can create new images by designing hand-drawn masks in a basic image editor (e.g. Paint 3D), leveraging the established colorbook. Our streamlined pipeline is capable of effortlessly producing images from these handcrafted sketches. See example on Figure 7.

Fig. 7: Rough hand painting

While the generated images might lack the intricate details seen in outputs from precise semantic masks, the enhanced user control and hands-on experience somewhat balance these drawbacks and mark a place for further improvements.

V. Discussion

A. Limitations and Future Work

While our model has been adeptly trained to generate objects such as cars, vegetation, buildings, and the sky at arbitrary locations, its performance falters when generating certain objects like human figures. This limitation can be attributed to the insufficient training samples of such classes. As illustrated in Figure 8, the model correctly identifies the spatial position for the human object, but fails to provide an accurate texture representation.

Fig. 8: Challenges in human figure generation

To further understand the model's limitations, we conducted an analysis of the class distribution within the training data. Figure 9 provides a visual summary of the pixel-wise occurrence of each class. Utilizing this information, the model's performance can potentially be enhanced by supplementing it with more training samples from underrepresented classes. Furthermore, the presented solution operated on a low-resolution, in alignment with the available computational resources.

Looking forward, there are multiple avenues for improvement. One promising direction is incorporating Vector Quantized (VQ) [14] encodings. As highlighted by [15], noising the encoded tokens and directly predicting the denoised versions (instead of noise) could provide both speed and scalability benefits. This approach could facilitate the integration of additional Attention blocks, presenting an exciting opportunity to refine the current architecture.

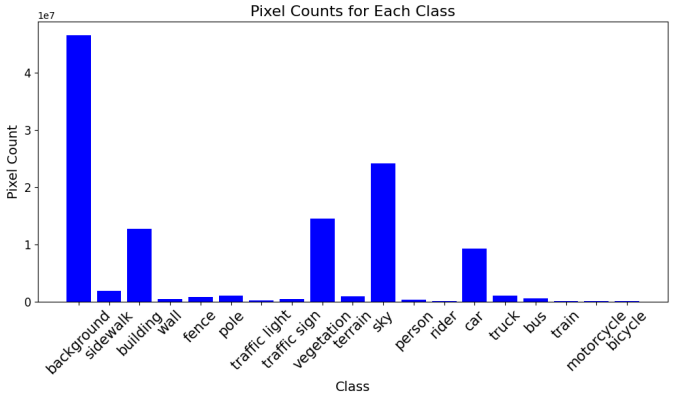

Fig. 9: Class distribution analysis in training data

Furthermore, for industrial application, upscaling the current method is inevitable. Besides modifying the compression rates, developing a dedicated upscaler diffusion model tailored to the applied dataset for this specific goal could substantially enhance the quality, yielding more captivating and visually appealing outcomes. Finally, utilizing textual prompts as an additional control mechanism would be beneficial, allowing for precise control not only over structure, but over the color attributes of the images as well.

VI. Summary and Conclusion

Throughout this work, we have integrated semantic segmentation mask guidance by modifying a standard U-Net-like architecture. Utilizing a DDPM-based [5] diffusion model, we successfully trained a semantic mask-guided model at a 128x128 resolution. We demonstrated that semantic segmentation masks provide clear control over scene generation. This versatility was showcased through various applications: mask-guided, modified mask-guided, and hand-painted drawing-guided generation. In addition to visual evidence, we conducted multiple metric-based evaluations, affirming the method's stability. While the resolution of the generated images may not rival that of larger, cutting-edge models, the degree of control surpasses what is typically achieved with mere text prompting. This finding encourages further enhancements of the model, especially in upscaling, which would make the method industrially applicable for data enrichment in the near future. The PyTorch Lightning [16] implementation of the training procedure is made publicly available for reproducibility and further research at https://github.com/kajc10/semseg-guided-diffusion.

Acknowledgment

The authors are grateful for the support of Continental Automotive Hungary Ltd. The work reported in this paper, carried out at BME, has been partly supported by the the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory and project no. TKP2021-NVA-02 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme.

References

[1] A.-S. Maerten and D. Soydaner, "From paintbrush to pixel: A review of deep neural networks in ai-generated art." arXiv preprint arXiv:2302.10913, 2023.

[2] Haq, Imran Ul. "An overview of deep learning in medical imaging." ArXiv abs/2202.08546, 2022.

[3] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, et al., "End to end learning for self-driving cars." arXiv preprint arXiv:1604.07316, 2016.

[4] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving" 2017.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models." Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.

[6] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion model in generative ai: A survey" arXiv preprint arXiv:2303.07909, 2023.

[7] Meng, Chenlin, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu and Stefano Ermon. "SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations." International Conference on Learning Representations (2021).

[8] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models." in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847, 2023

[9] Weilun Wang and Jianmin Bao and Wengang Zhou and Dongdong Chen and Dong Chen and Lu Yuan and Houqiang Li, "Semantic Image Synthesis via Diffusion Models.",2022

[10] J. Ho and T. Salimans, "Classifier-Free Diffusion Guidance." NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2021.

[11] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization." 7th International Conference on Learning Representations ICLR, 2019.

[12] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov and R. Girdhar, "Masked-attention Mask Transformer for Universal Image Segmentation." IEEE/CVF Computer Vision and Pattern Recognition Conference CVPR, 2022

[13] L.C. Chen, G. Papandreou, F. Schroff, H. Adam, "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587, 2017.

[14] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations." Advances in neural information processing systems, vol. 30, 2017.

[15] D. Rampas, P. Pernias, and M. Aubreville, "A novel sampling scheme for text- and image-conditional image synthesis in quantized latent spaces." 2023.

[16] William Falcon and The PyTorch Lightning team, GitHub. Note: https://github.com/Lightning-AI/lightning, 2019.

VII. Appendix

Fig. 10: More examples of test masks and corresponding generations