Supplementary Materials: "Seg-diffusion: Text-to-Image Diffusion Model for Open-Vocabulary Semantic Segmentation"

Anonymous Authors

In this supplementary material, we provide additional implementation details and experimental results. Our code is publicly available at: https://github.com/QuantumScriptHub/Seg-diffusion.

1 IMPLEMENTATION DETAILS

Multi-resolution noise. Prior research has investigated variations from the classical DDPM approaches, introducing elements like atypical noise distributions or shortcuts that diverge from the Markovian sequence. The framework we introduce, along with the detailed fine-tuning procedure, accommodates alterations to the noise sequence during the fine-tuning phase. We discovered that a synergistic blend of noise applied at multiple resolutions, coupled with a gradually decreasing schedule, accelerates convergence and significantly enhances outcomes compared to the conventional DDPM approach. This layered noise approach involves layering a series of Gaussian noise patterns at varying scales, each scaled up to match the resolution required for the input of U-Net. Our suggested graduated schedule smoothly transitions from this layered, multi-resolution noise at $t = T$ to the typical Gaussian noise at $t = 0$. Furthermore, the gradual reduction of multi-resolution noise contributes to an additional improvement. We have also observed that training with multi-resolution noise results in more consistent predictions, even when different initial noise is used during inference. The annealing process further enhances this consistency.

Test-time ensembling. The probabilistic nature of the inference pipeline results in diverse predictions, contingent upon the initialization noise present in $\mathbf{z}_T^{\mathbf{y}}$. Leveraging this, we introduce a test-time ensemble approach capable of amalgamating numerous inference iterations on the same input For every input sample, there is the option to perform inference up to N times. This approach provides a versatile balance between computational efficiency and predictive accuracy, determined by the chosen value of N. We have observed a consistent improvement as the ensemble size increases. However, this improvement begins to diminish after reaching 10 predictions per sample.

Latent diffusion denoising. We encode the input image into the latent space, initialize mask latent as standard Gaussian noise, and progressively denoise it with the same schedule as during fine-tuning. We empirically find that initializing with standard Gaussian noise gives better results than with multi-resolution noise, although the model is trained on the latter. We follow DDIM's approach to perform non-Markovian sampling with re-spaced steps for accelerated inference. The final segmentation map is decoded from the latent code using the VAE decoder and postprocessed by averaging channels.

Denoising steps. We assess the impact of re-spaced inference denoising steps guided by the DDIM scheduler. Despite being trained with 1000 DDPM steps, opting for 50 steps proves adequate for accurate inference results. As anticipated, employing more denoising steps leads to superior outcomes. We note that the point of diminishing returns with additional denoising steps varies across datasets but consistently remains below 10 steps. This suggests that denoising steps can be further reduced to 10 or fewer for increased efficiency while maintaining comparable performance. Interestingly, this threshold is lower than the typical requirement of 50 steps for diffusion-based image generators.

Training noise. We investigate the impact of three types of noise during the training phase. Training with multi-resolution noise significantly improves the depth prediction accuracy over using standard Gaussian noise. Furthermore, the gradual annealing of multi-resolution noise yields an additional improvement. We also noticed that training with multi-resolution noise leads to more consistent predictions given different initial noise at inference time and annealing further enhances this consistency.

Number of denoising steps. We evaluate the effect of the re-spaced inference denoising steps driven by the DDIM scheduler. Although trained with 1000 DDPM steps, the choice of 50 steps is sufficient to produce accurate results during inference. As expected, we obtain better results when using more denoising steps. We observe that the elbow point of marginal returns given more denoising steps depends on the dataset but is always under 10 steps. This implies that one can further reduce the denoising steps to 10 or even less to gain efficiency while keeping comparable performance. Interestingly, this threshold is smaller than what is usually required for diffusion-based image generators, *ie*, 50 steps.

2 EXPERIMENTAL RESULTS

We will now present additional comparative experiments. From left to right, they are: Original, OVSeg, SAN, Ours, and GT. Clearly, compared to other methods, our Seg-diffusion has achieved excellent results. For more implementation details, please read and clone our GitHub repository.


clouds
skyscraper
fog
tree
bush
banner
fence
wall-other
person
wall-concrete
skateboard
sky-other
pavement
building-other
floor-stone
clock
parking meter
desk-stuff
car
metal
grass
roof
road
clothes
bottle
motorcycle
horse
bush
building-other
person
backpack
cell phone
tie
orange
snowboard
toothbrush
traffic light
handbag
light
skateboard
ground-other
mountain
house
sports ball
baseball bat
gravel
ceiling-other
cupboard
wall-panel
refrigerator
counter
cabinet
chair
dining table
floor-other
microwave
toaster
oven
sink
wall-concrete
wall-tile
window-other