Multi-Class Segmentation from Aerial Views using Recursive Noise Diffusion: Supplementary Material

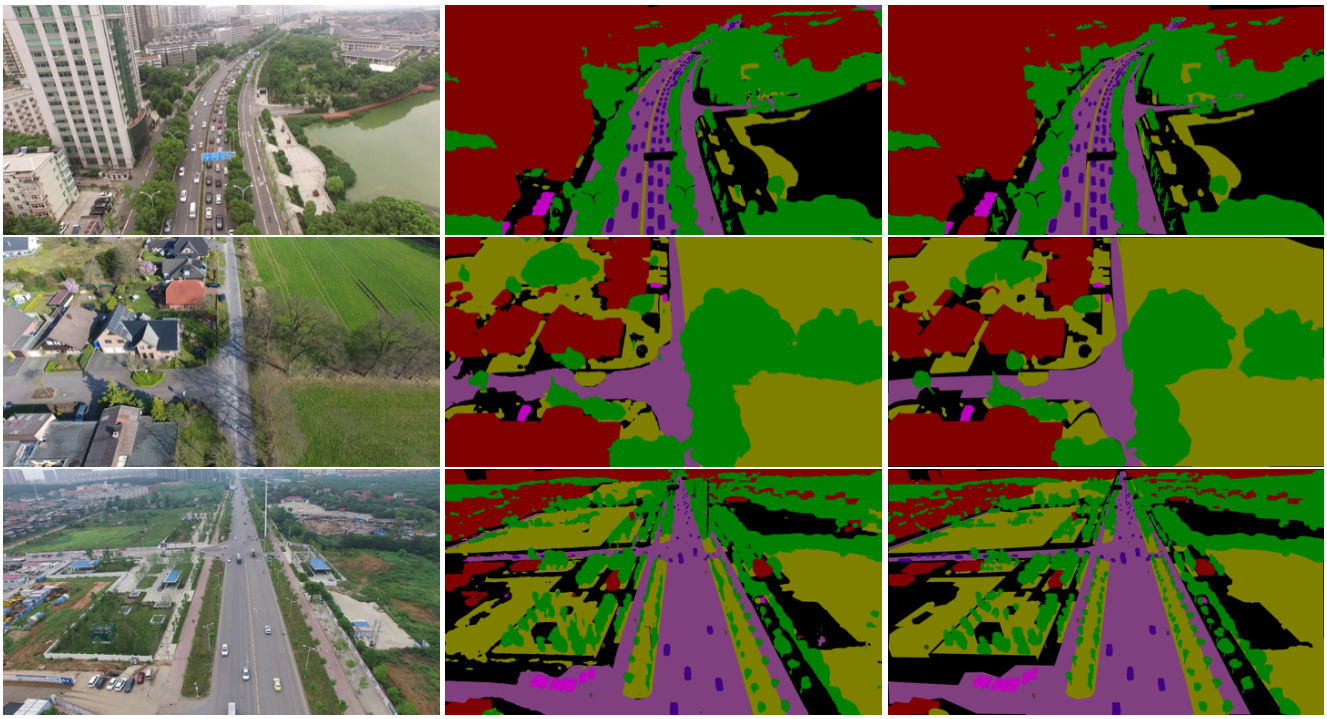
Benedikt Kolbeinsson
Imperial College London
bk915@imperial.ac.uk

Krystian Mikolajczyk
Imperial College London
k.mikolajczyk@imperial.ac.uk

A. Qualitative experiments

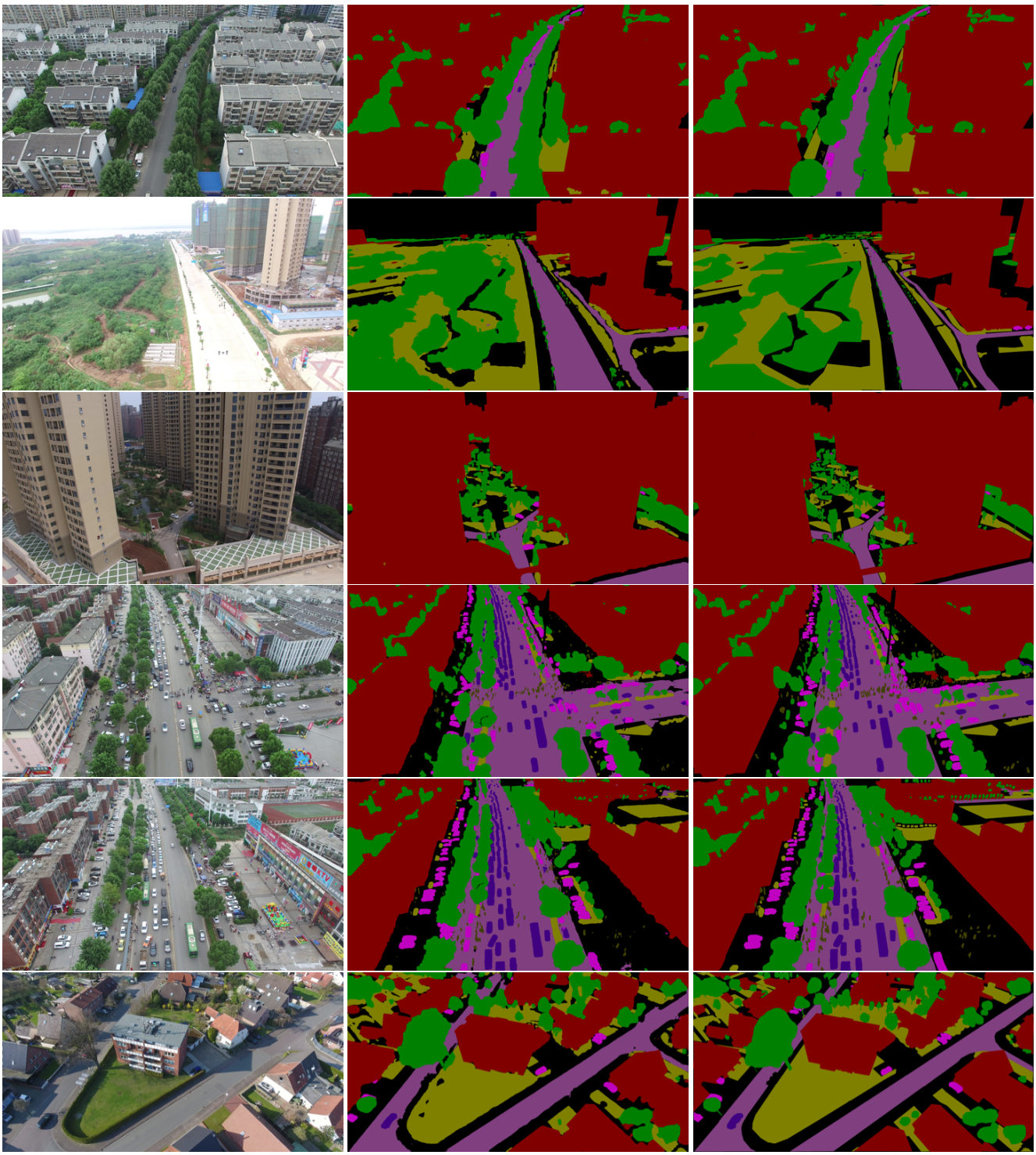
In this section, we provide additional results and qualitative comparisons to further support the effectiveness and robustness of our approach. Figures 6 and 7 show qualitative results on the UAVid [3] validation dataset, using our multi-class segmentation approach. Our method shows remarkable ability in segmenting moving cars and static cars, given only a single static input. This indicates our method uses contextual cues found in the surrounding scene to determine if a car is parked or moving. Our method can in some cases produce better segmentation than the ground truth, as can be seen in Figure 6.

Figure 8 shows additional qualitative results on Vaihingen Buildings [1]. Notably, our method demonstrates a deep understanding of the scene and is able to isolate and segment the primary center building with precision and accuracy, despite the presence of numerous other buildings and structures in the surrounding area. Images are best viewed on a computer and zoomed-in.



(a) Image (b) Ours (c) Ground truth

Figure 6. **Qualitative results on UAVid [3] validation**. Our method can detect small details such as street lights, tree branches and people. The ground truth labels are course and often missing, as can be seen in the first example where the street lights in the center of the image are missing and the tall telephone tower in the bottom example is missing from the ground truth.



(a) Image (b) Ours (c) Ground truth

Figure 7. **Additional qualitative results on UAVid [3] validation**. It is notable how our method can correctly classify static and moving cars given only a single static input image.

Figure 8. **Additional qualitative results on Vaihingen Buildings [1].** For each row group, top row: input image, middle row: our method, bottom row: ground truth. Our method struggles most when there is a building extension with a different colored roof as can be seen in the bottom row, third from the left.

B. Architecture

Our model is composed of four key modules: a time step head, an image encoder head, a diffused segmentation encoder head, and the primary UNet-like encoder-decoder. To integrate the time step into the model, the time step head transforms it into a sinusoidal positional embedding, inspired by the positional embeddings utilized in Vaswani *et al*. [6]. The image and segmentation heads have the same structure, each including two ResNetBlocks (as shown in Figure 9, but notably without time embeddings). The sum of the outputs, of the image and segmentation heads, are passed to the encoder-decoder. Our encoder-decoder module takes inspiration from Efficient U-Net [4]. The architecture of our encoder-decoder is shown in Figure 10, which incorporates time step embeddings with each ResNetBlock. Additionally, our model leverages Efficient Attention [5], a type of attention mechanism with linear complexity.

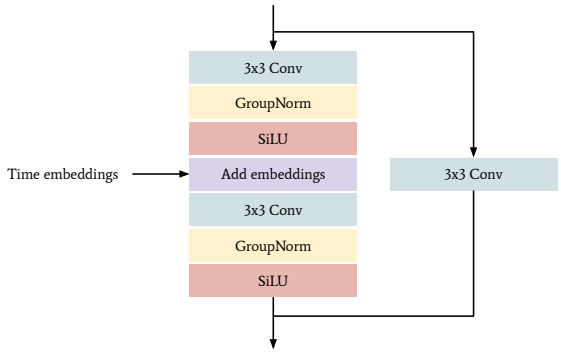


Figure 9. **Diagram of our ResNetBlock**, consisting of a residual connection [2], a core building block for the model.

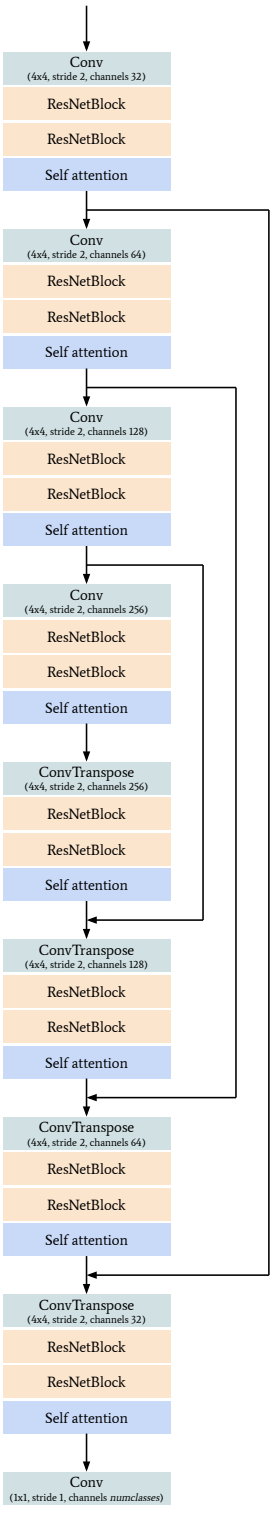


Figure 10. **Schematic of our encoder-decoder**. The self attention block uses Efficient Attention [5]. The details of the ResNetBlocks are shown in Figure 9.

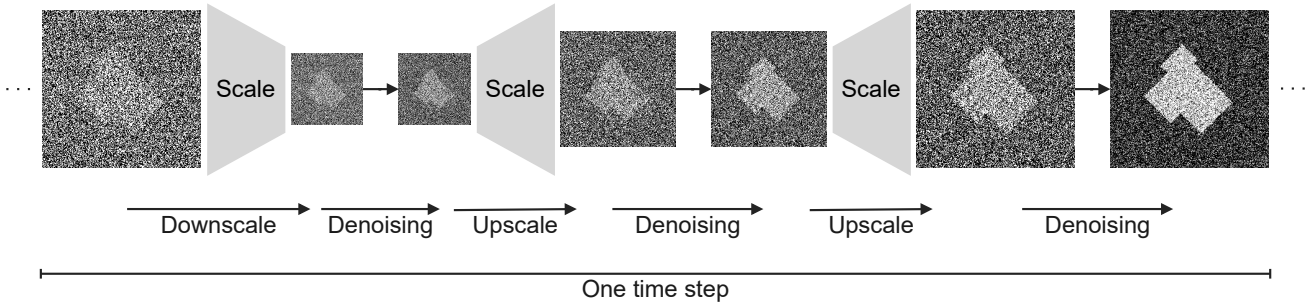


Figure 11. **Variation on the scaling schedule.** At each time step, the input is resized to its smallest scale and the diffused segmentation is denoised at this lower scale. The denoising process is then repeated iteratively as the segmentation is upscaled until it reaches its original scale.

C. Multi-scale schedule variation

For the Vaihingen Buildings dataset we found a modification to the scaling schedule worked better. This modification involves denoising the image at each scale for each time step and can be seen in Figure 11. For each time step, the input is downscaled to the smallest scale (if there are multiple scales) and the diffused segmentation is denoised at this smaller scale. Then the segmentation is upscaled and denoised, repeatedly until the original scale is reached. We use bilinear interpolation for both downscaling and upscaling. Training with this scaling schedule is shown in Algorithm 2. We found this scaling schedule worked better on Vaihingen Buildings than the linear scale scheduling, which we used for UAVid.

Algorithm 2: Training with hierarchical scales

Input: $\mathbf{x} \in \mathbb{R}^{W \times H \times 3}$, RGB image
Input: $\bar{\mathbf{s}} \in \mathbb{R}^{W \times H \times classes}$, segmentation labels
Parameters: $T \in \mathbb{Z}^1$, number of time steps
Parameters: $M \in \mathbb{Z}^1$, number of scales

$\hat{\mathbf{s}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
  for $m = M, ..., 1$ **do**
    Resize $\hat{\mathbf{s}}_t$ to size $(\frac{W}{2^{m-1}} \times \frac{H}{2^{m-1}} \times \text{classes})$
    Resize $\mathbf{x}$ to size $(\frac{W}{2^{m-1}} \times \frac{H}{2^{m-1}} \times 3)$
    $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{s}'_t \leftarrow \hat{\mathbf{s}}_t + \mathbf{z}_t \times \frac{t}{T}$ // diffuse
    $\hat{\mathbf{s}}_{t-1} \leftarrow \mathbf{s}'_t - \epsilon_\theta(\mathbf{s}'_t, \mathbf{x}, t)$ // denoise
    $l \leftarrow \|\epsilon_\theta(\mathbf{s}'_t, \mathbf{x}, t) - (\mathbf{s}'_t - \bar{\mathbf{s}})\|^2$
    Update ϵ_θ w.r.t. l
  end
end

References

[1] Michael Cramer. The dgpf-test on digital airborne camera evaluation overview and test design. *Photogrammetrie-Fernerkundung-Geoinformation*, pages 73–82, 2010. 1, 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[3] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108–119, 2020. 1, 2

[4] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 4

[5] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021. 4

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4