DocWaveDiff: A Predict-and-Refine approach for Document Image Enhancement with Wavelet U-Nets and Diffusion models — Supplementary material

A. Analysis of memory consumption and execution

GPU memory Vs. diffusion steps T. Our measurements indicate that the memory peak remains essentially constant as T varies ($\approx$ 4.92 GB for $T = 10$–90, 5.06 GB at $T = 100$), see Figure 1. This behavior is expected: during sampling, the diffusion steps are sequential and reuse the same buffers; at each step, the same workspace (UNet activations, attention and wavelet) is allocated on fixed-size inputs with a constant batch size of 64. Therefore, the number of diffusion steps has no effect on memory consumption, which is dominated by model weights, input size, and batch size.

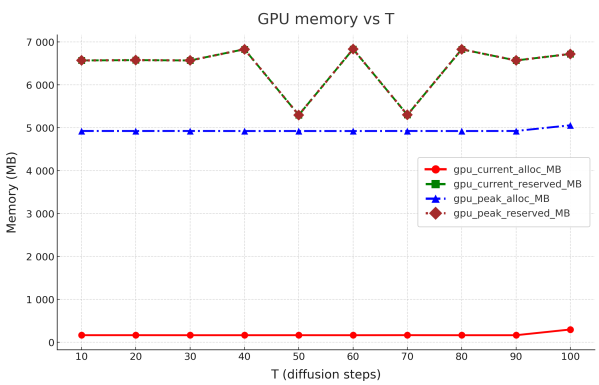

Figure 1. **GPU memory vs T (diffusion steps).** x-axis: T; y-axis: memory (MB). Curves show current/peak *allocated* and current/peak *reserved* memory (see legend).

Time execution Vs. diffusion steps T. In order to quantify the impact of inference times, the inference times of a patch were measured, varying the number of diffusion steps. The sampling time increases almost linearly with steps, Figure 2, as predicted by diffusion methods.

B. Model architecture

This section of the additional material contains the complete DocWaveDiff schema. In addition to the model, this complete version includes the schema of attention layers, res-blocks, and time embeddings, Figure 3.

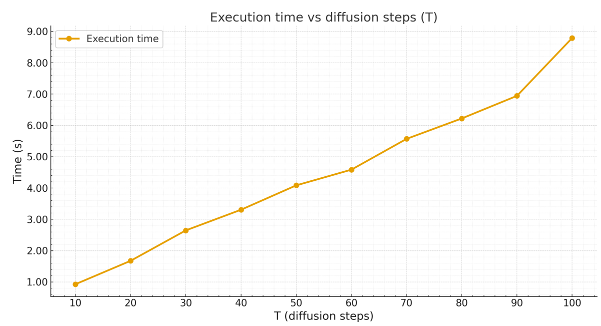

Figure 2. **Time execution vs T (diffusion steps).** x-axis: T; y-axis: second.

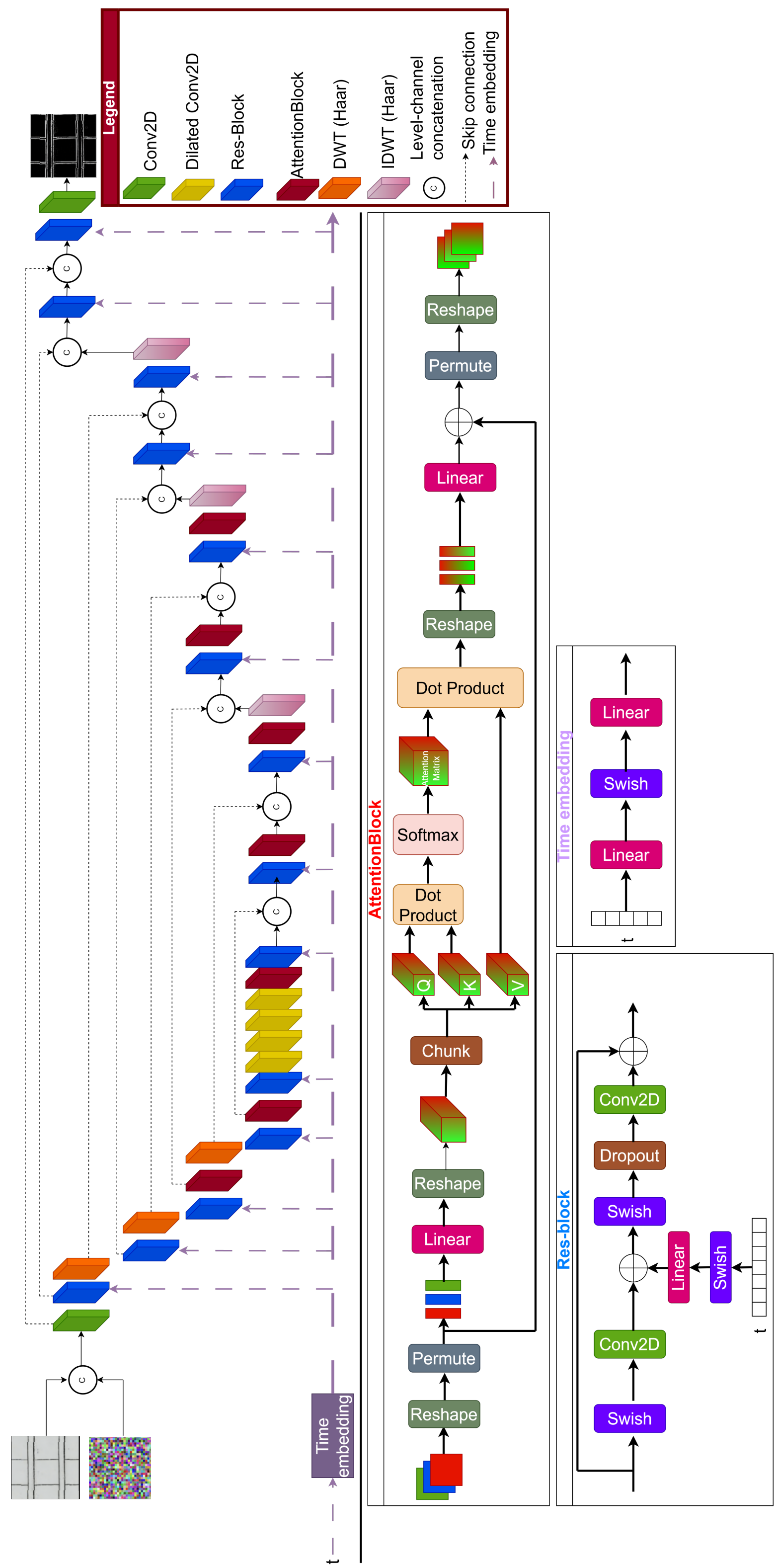

Figure 3. The proposed Figure highlights the wavelet U-Net architecture for the Refiner Denoiser. The design is accompanied by a legend illustrating the blocks that make up the network. For completeness, the design of the ResBlock, Attention Block, and time embedding are shown next to the U-Net architecture.