DICE Rebuttal

A. Introduction

We would like to thank the reviewers for their meaningful and constructive comments on the paper, and wholely accept them as limitations that need to be addressed for a more refined submission.

Nonetheless, we would primarily like to offer some clarification on several key points that may hopefully at least partially address these concerns.

B. Response

B.1. Performance (sxgP, 8jwP, jW43)

The reviewers all validity critiqued the provided high FID scores. This major issue stems from the lack of compute available to properly validate the results. However, we attempted to address this concern by seeding the generated images such that the latent noise was the same for all 500 samples. Understandably this technique doesn't provide an in-depth analysis of performance, but still we hope to show that it is a valid method to produce comparable results between the models tested.

B.2. Motivation (sxgP, jW43)

Here, the reviewers questioned the motivation behind why each level's output is treated as a noise prediction. This process is driven by the how the diffusion proccess operates to estimate noise from the given latent input. Our major contribution was managing each level such that each lower stage 'assists' the upper ones by preliminarily denoising their inputs. This was identified by attempting to optimise the utilisation of the lower stage which sees the most global scope, which would otherwise inadvertently introduce noise when upsampled in conventional cascade models, as shown in Figures 1 to 3.

B.3. Artifacting (sxgP, jW43)

We would like to apologize as there is a minor mistake on our part in our report, as Figures 2 to 4 from the original paper were created using nearest approximation rather than linear interpolation as otherwise suggested in section 3.1.

For the sake of completeness, we provide the results of both interpolation methods to show that both methods would still introduce noise for the same provided reason.

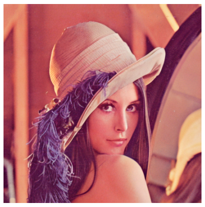

Figure 1.
1x nearest residual

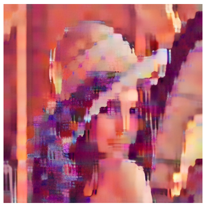

Figure 2.
2x nearest residual

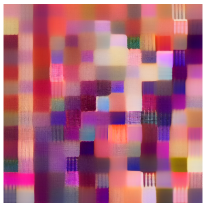

Figure 3.
4x nearest residual

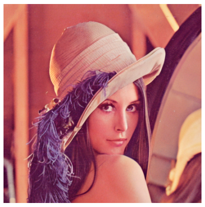

Figure 4.
1x bilinear residual

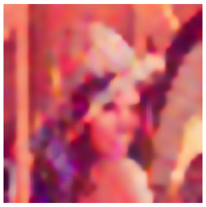

Figure 5.
2x bilinear residual

Figure 6.
4x bilinear residual

The reviewers also commented for clarification for section 3.1. The key argument that we make here is that the upsampled residual connections introduce noise as a fundamental result of interpolating tokens. In the case of DiMR, while this doesn't result in a failure of training, we argue that this hampers the models' ability to efficenctly process these inputs.

Instead, we attempt to avoid this issue entirely by decoding what would be the residual connection into the form of a noise prediction which is optimised by the loss function. From here, we can leverage these predictions as they are partial solutions within the diffusion process itself, which conveniently allows us to maximise the utilisation of each layer. As stated before, this is done by denoising the inputs to the upper stages, but they can also be added together to predict the final noise output, thus completing the cycle.

We concor that this solution is fairly abstract as it attempts to solve multiple issues within cascading models at once, but we hope to at least provide some reasoning behind our solution and why it is structured the way that it is.