Progressive Artwork Outpainting via Latent Diffusion Models Supplementary Materials

Dae-Young Song[1], Jung-Jae Yu[1], Donghyeon Cho[2]*
[1]Electronics and Telecommunications Research Institute, Daejeon, South Korea
[2]Department of Computer Science, Hanyang University, Seoul, South Korea
{eadyoung,jungjae}@etri.re.kr, doncho@hanyang.ac.kr

In this supplementary material, additional items are provided as follows:

- Detailed description of data augmentation tricks.
- Additional implementation details and default prompts.
- Applications on real image dataset (landscape).
- Discussion on the aim of the proposed methods.
- Additional qualitative results and blurriness visualization.

1. Data Augmentation

Algorithm 1 Obtaining coordinates of random patches

Input: image size H, W, window size K
Parameter: variation of window $q \leq 1$, spare pixel u
Output: size k, coordinates x, y

Random window length, $k \leftarrow$ randint(qK, K).
Erased height $r_h \leftarrow \frac{rH}{2}$.
Erased width $r_w \leftarrow \frac{rW}{2}$.
Height spare pixels, $u_h \leftarrow u$ if $r_h \geq k$ else 0.
Width spare pixels, $u_w \leftarrow u$ if $r_w \geq k$ else 0.
Random direction flag $d \in \{$up, down, left, right$\}$.
if d == up **then**
    $y \leftarrow$ randint(max$(0, r_h - k) + u_h, r_h - u_h)$.
    $x \leftarrow$ randint(max$(0, r_w - k) + u_w, W -$ max$(r_w, k) - u_w)$.
else if d == down **then**
    $y \leftarrow$ randint($H - (r_h + k) + u_h, H -$ max$(r_h, k) - u_h)$.
    $x \leftarrow$ randint(max$(0, r_w - k) + u_w, W -$ max$(r_w, k) - u_w)$.
else if d == left **then**
    $y \leftarrow$ randint(max$(0, r_h - k) + u_h, H -$ max$(r_h, k) - u_h)$.
    $x \leftarrow$ randint(max$(0, r_w - k) + u_w, r_w - u_w)$.
else if d == right **then**
    $y \leftarrow$ randint(max$(0, r_h - k) + u_h, H -$ max$(r_h, k) - u_h)$.
    $x \leftarrow$ randint($W - (r_w + k) + u_w, W -$ max$(r_w, k) - u_w)$.
end if
return k, x, y

In this section, we describe the detailed techniques for training data augmentation in the training strategies section of the main paper. For the given image $I^e \in \mathbb{R}^{H \times W \times 3}$ in Figure 6 of the main paper, we first erase inner area of the original by $\frac{rH}{2}$

in height and $\frac{rW}{2}$ in width. The deletion ratio r, variation parameter q, and spare pixels u are hyperparameters that determine how much of the original image is reduced, the crop size, and the extent of overlap with known pixels, respectively. They influence the difficulty of training; we set them to 0.5, 0.9, and 4, respectively. After that, we randomly attach some patches to the reduced image I_C to simulate an arbitrary intermediate step image I_τ^e at τ. The top-left coordinates of each random patch can be obtained using the Algorithm 1. Specifically, under the Figure 7-(a) scenario in the main paper, the final I_C and

Algorithm 2 Random Attachments on I_C

Input: given image I^e, initial I_C, K
Parameter: maximum attaching numbers N, escape ratio e
Output: final I_C, M_{global}

Random attaching number $n \leftarrow \text{randint}(0, N)$.
$i \leftarrow 0$.
$I^e \in \mathbb{R}^{H \times W \times 3}$.
$M_{global} \in \mathbb{R}^{H \times W} \leftarrow 1$.
while $i <$ range(n) **do**
  $k, x, y \leftarrow$ Algorithm 1(H, W, K).
  $I_C[y : y + k, x : x + k] \leftarrow I^e[y : y + k, x : x + k]$.
  $M_{global}[y : y + k, x : x + k] \leftarrow 0$.
  if mean$(M_{global}) \leq e$ **then**
    break.
  end if
end while
return final I_C, M_{global}

the corresponding mask M_{global} are obtained by applying the random attachments as described in Algorithm 2. K is 512 in the frozen stable diffusion (SD) inpainting model. N and e are a maximum number of attachments and an escape ratio, and we set $N = 32$ and $e = 0.15$, respectively. During Algorithm 2, the attachment process can result in an overabundance of known pixels, leaving insufficient region for outpainting. To address this case, e is introduced as a termination criterion for Algorithm 2. This ensures that the process halts before attempting to fulfill the random attaching number n. Consequently, they are also hyperparameters that determine the training difficulty. After Algorithm 2, the position of the local window needs

Algorithm 3 Positioning of the local window

Input: final I_C, M_{global}, K
Parameter: termination parameter d
Output: top-left coordinates of local window l_x, l_y

$I_C \in \mathbb{R}^{H \times W \times 3}$.
while True **do**
  $l_x \leftarrow \text{randint}(0, W - K)$.
  $l_y \leftarrow \text{randint}(0, H - K)$.
  $M \leftarrow M_{global}[l_y : l_y + K, l_x : l_x + K]$.
  if mean$(M) > d$ **then**
    break.
  end if
end while
return l_x, l_y

to be selected within the final image I_C and M_{global} to acquire I, M, and M_C. Thus, Algorithm 3 is utilized to obtain the top-left coordinates l_x and l_y, which indicate the positions of M_C, I, and M. d is a decision parameter that terminates Algorithm 3 if the white portion (generation area) of M exceeds this parameter, and we set it to 0.05. The purpose of d is to prevent cases where the position is determined in such a way that there are no regions left to be generated during the augmentation process.

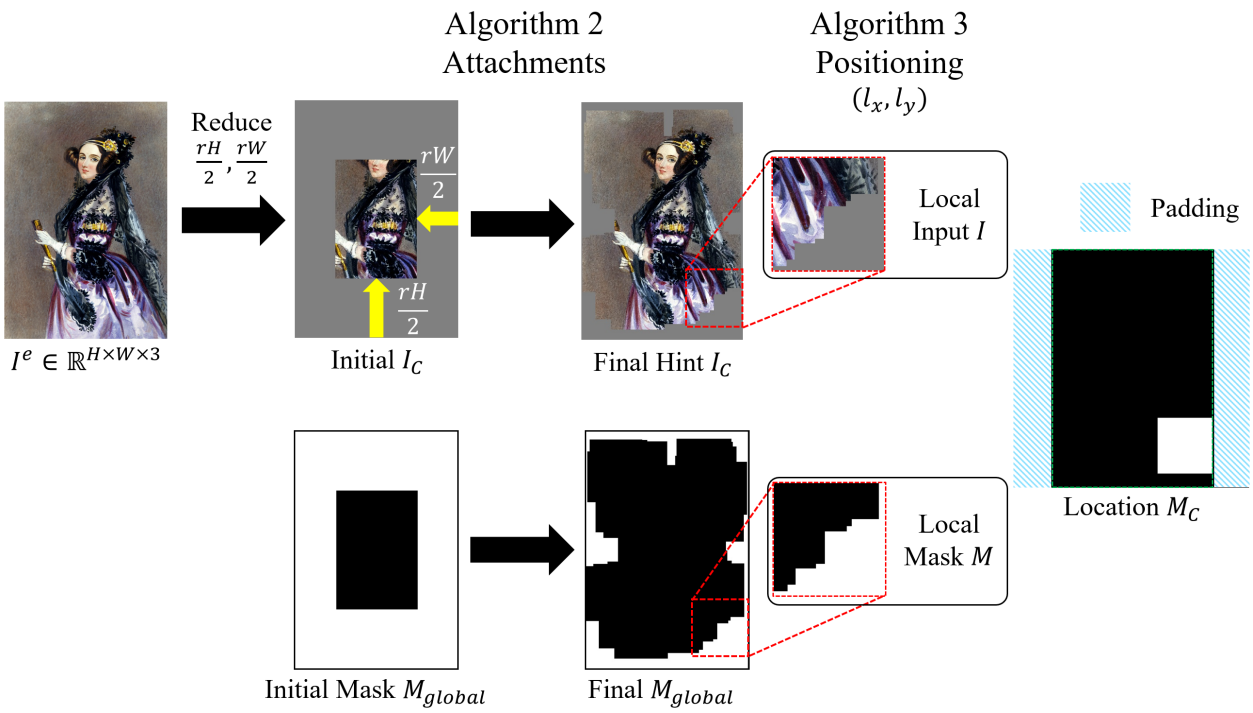

Figure 1. An overall workflow of data augmentation.

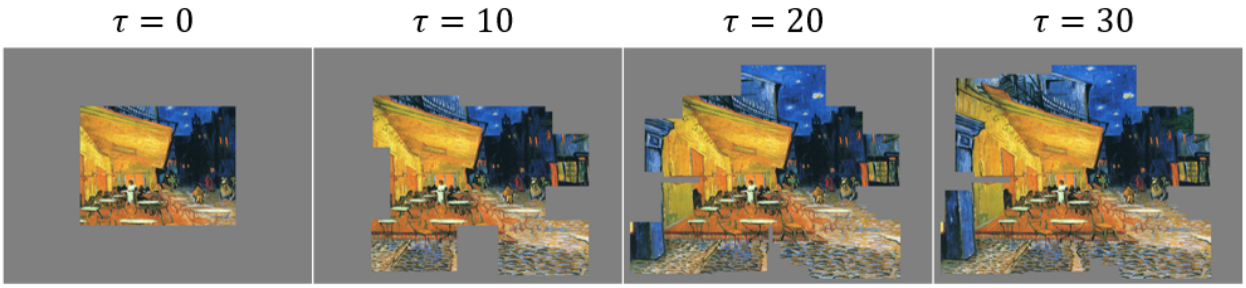

Figure 2. Visualization of augmented images across different τ values. Without the proposed augmentation method, the progressive outpainting models are exposed to $\tau = 0$ data during training.

The overall workflow corresponding to Algorithm 2 and Algorithm 3 is illustrated in Figure 1. As Illustrated in Figure 2, the proposed augmentation method prevents the model from being exposed solely to $\tau = 0$ during training, thereby mitigating overfitting of CPMs. To this end, we randomly sample an n to simulate various τ steps as described in Algorithm 2.

2. Additional Implementation Details

In this section, we describe the local condition c_L and compare the total number of parameters required for inference. In progressive outpainting, the global condition c_G obtained from pre-trained model like BLIP [9] is unsuitable for input into the frozen SD branch, which handles localized generation. Instead, c_L is supplied for the SD branch. We empirically select keywords that can be used broadly in terms of qualitative aspects.

Default Prompts we use *"harmonized painting, high resolution, best quality, high quality, harmonized simple background"* as a default local condition c_L. In addition, a negative prompt can be utilized for classifier-free guidance [7]. Therefore, we use *"ugly, nsfw, (text:1.5), (copyright:1.25), (blurry:1.5), worst quality, watermark, signature, logo"* as a default negative prompt for the frozen SD branch. A prompt in the format "$(k : w)$" indicates the use of prompt reweighting [5]. For example, if w is 2, the prompt is weighted twice as strongly, and if w is 0.5, it is weighted at half strength. At the same time,

we can optionally inject a negative global condition c_G as follows: *"a painting in (a frame), (collage drawing), (divided painting:1.25), crop, text"*.

The Number of Model Parameters To evaluate the computational complexity of the proposed methods, particularly the

SD Inpainting	Ours (ControlNet)	Ours (Fusion)
1,066,249,707	1,427,543,371	1,449,137,963
-	# of parameters of CPM	
-	361,293,664	382,888,256

Table 1. The total number of parameters for each model. The total parameters count include the *variational autoencoder* for compression into the latent space, the *CLIP* model for text-to-image conditioning, the *SD denoising U-Net*, and the *CPM*.

newly introduced fusion-based composition planning module (CPM), we report the total number of parameters required for outpainting inference. As shown in Table 1, the parameter count of the fusion-based CPM method is slightly larger than that of the ControlNet-based [13] CPM. Note that the fusion-based CPM is trained from scratch for 50 epochs, while the ControlNet-based CPM is trained through transfer learning by duplicating the pre-trained encoder of the SD model.

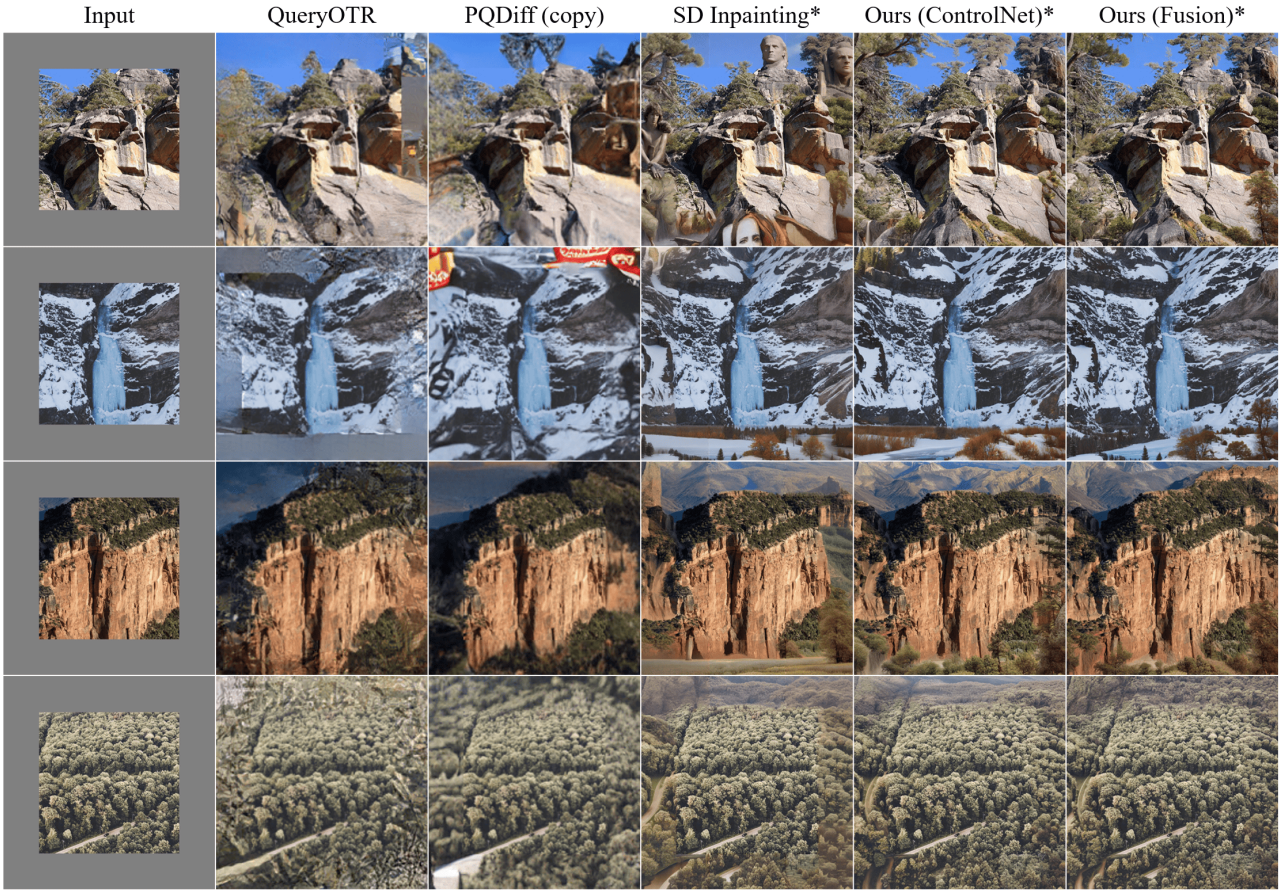

Figure 3. Qualitative results of the LHQ dataset based on the scenario of Figure 7-(a) in the main paper. An "*" indicates that the method based on the progressive outpainting, whereas the others generate in all directions at once and are supported with super-resoultion. Same seed value is used for the SD-based models.

3. Applications on Real Image Dataset

We sampled 1,999 natural landscape images from the LHQ [2] dataset. Although these images contain repetitive patterns (e.g., forests) and are less sensitive to composition than artworks, they present a challenging setting to evaluate the generalization ability of the proposed methods. For this experiment, the case in Figure 7-(a) in the main paper is applied with $r = 0.3$ as illustrated in Figure 3. As shown in Table 2, the existing methods suffer from limited fidelity of local details, achieving unsatisfactory patch FID scores.

Metric	CLIP-aes↑	pFID@256↓	pFID@512↓
QueryOTR	4.236	71.08	38.21
PQDiff (gen)	3.546	99.25	61.59
PQDiff (copy)	4.005	72.16	40.60
SD Inpainting	5.187	31.72	23.93
Ours (ControlNet)	5.486	20.99	14.56
Ours (Fusion)	**5.541**	**19.65**	**14.35**

Table 2. Evaluation on 1,999 random samples from LHQ dataset. Note that the results of QueryOTR and PQDiff are resized to the original aspect ratio and then processed with SR. The "copy" of the PQDiff method refers to a copy-and-paste post-processing approach, where the original input is merged into the outpainting results, whereas "gen" does not. **bold**: best.

4. Discussion on the Aim of the Proposed Methods

In order to clarify the distinct intent of our study, this section compares our method with representative reference-based image generation approaches. Existing methods usually feed the reference image into the generator and directly inject its prominent content into the output. On the other hand, our methods treat the reference as a compositional cue and infer semantically consistent content from the global scene, even when the reference image itself lacks the explicit corresponding content. This context-aware inference clearly distinguishes our work from previous research.

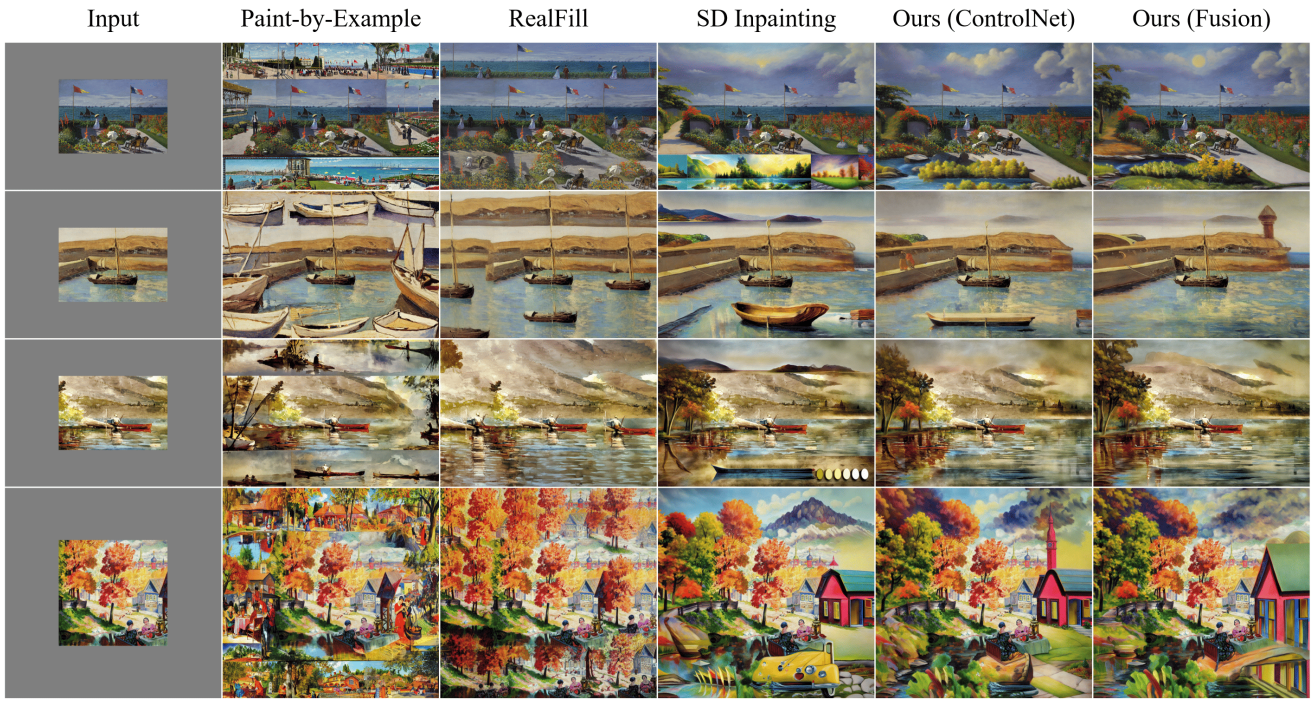

Figure 4. Qualitative results of the WikiArt dataset based on the scenario of Figure 7-(b) in the main paper with $r = 1$. All methods are based on the progressive outpainting. RealFill is trained for 2,000 steps with LoRA [8].

Figure 5. Qualitative results of the WikiArt dataset based on the scenario of Figure 7-(b) in the main paper with $r = 1$. All methods are based on the progressive outpainting. RealFill is trained for 2,000 steps with LoRA.

To conduct this experiment, we select Paint-by-Example [1] and RealFill [3], two reference-based image generation methods. The existing reference-based methods are designed to depend heavily on the reference image, so they often replicate its content in the generated images. This tendency is evident in Figure 4 and Figure 5. Although these methods belong to the category of reference-based generation, they are mainly effective for inpainting, where users deliberately insert specific elements, but they are less appropriate for outpainting. Their limitation is tolerable when the scene contains repetitive patterns such as the mountains or lake in Figure 5, yet it becomes pronounced for structured scenes in which composition and spatial

arrangement are crucial. In contrast, our approach places appropriate content by leveraging the global context of the reference image, and it does so without labor-intensive conditions, such as explicit object reference images or user prompts. As a result, artists who work with limited VRAM can split and extend high-resolution canvases with fewer trial-and-error iterations and reduced manual overhead.

5. Additional Results and Visualization

In this section, we present additional qualitative results that could not be included in the figures of the main paper. First, we present the blurriness map of Figure 9 of the main paper. The SD inpainting model, the ControlNet-based CPM model, and the fusion-based CPM model performed the progressive outpainting, while the others did not. Figure 6 shows the level of blurriness in the Figure 9, where brighter regions indicate higher levels of blur. Notably, we observe significant disparities in the blur intensity between the inputs and generated regions produced by the Firefly method, whereas the other methods maintain consistent levels. In addition, we present the qualitative results corresponding to the Table 1 and 2 in the main paper. Initially, Figure 7 illustrates the results of the IconArt experiment corresponding to the Table 1. Since FID [6], CLIP text score [10], and CLIP aesthetic score [11] are computed after resizing the results to a smaller size for input into the pre-trained networks, these quantitative results are included in the main paper. However, as shown in Figure 7, the existing methods with super-resolution generally suffer from severe sharpness degradation. Therefore, the blur measurement results are excluded from the Table 1 of the main paper to ensure a fair comparison.

Secondly, we report additional qualitative results of the experiment corresponding to the Table 2 in the main paper. As shown in Figure 8, the blur intensities of the results generated by the proposed methods appear generally darker compared to those of Firefly. Furthermore, as observed in Figure 6, Figure 7, and Figure 8, the proposed methods minimize the blurriness disparity between the input and the generated regions.

Lastly, we provide the qualitative results of the performance retention experiment for different ratios r of the task, as described in the ablation study and Table 3 of the main paper. As shown in Figure 9, the model without the CPM becomes increasingly prone to generating contextually irrelevant results and exhibits greater susceptibility to blur as the r increases. For example, it produces awkward mountains at the top of landscape paintings or duplicated faces or figures in portraits. In contrast, as illustrated in Figure 10 and Figure 11, outpainting models with the CPM produce natural landscapes that harmonize with the surrounding content. Additionally, their blur maps appear darker compared to Figure 9, reflecting improved sharpness. Notably, in vertically elongated portraits, the duplication of faces or figures is significantly mitigated, resulting in more coherent and visually plausible outpainting outcomes.

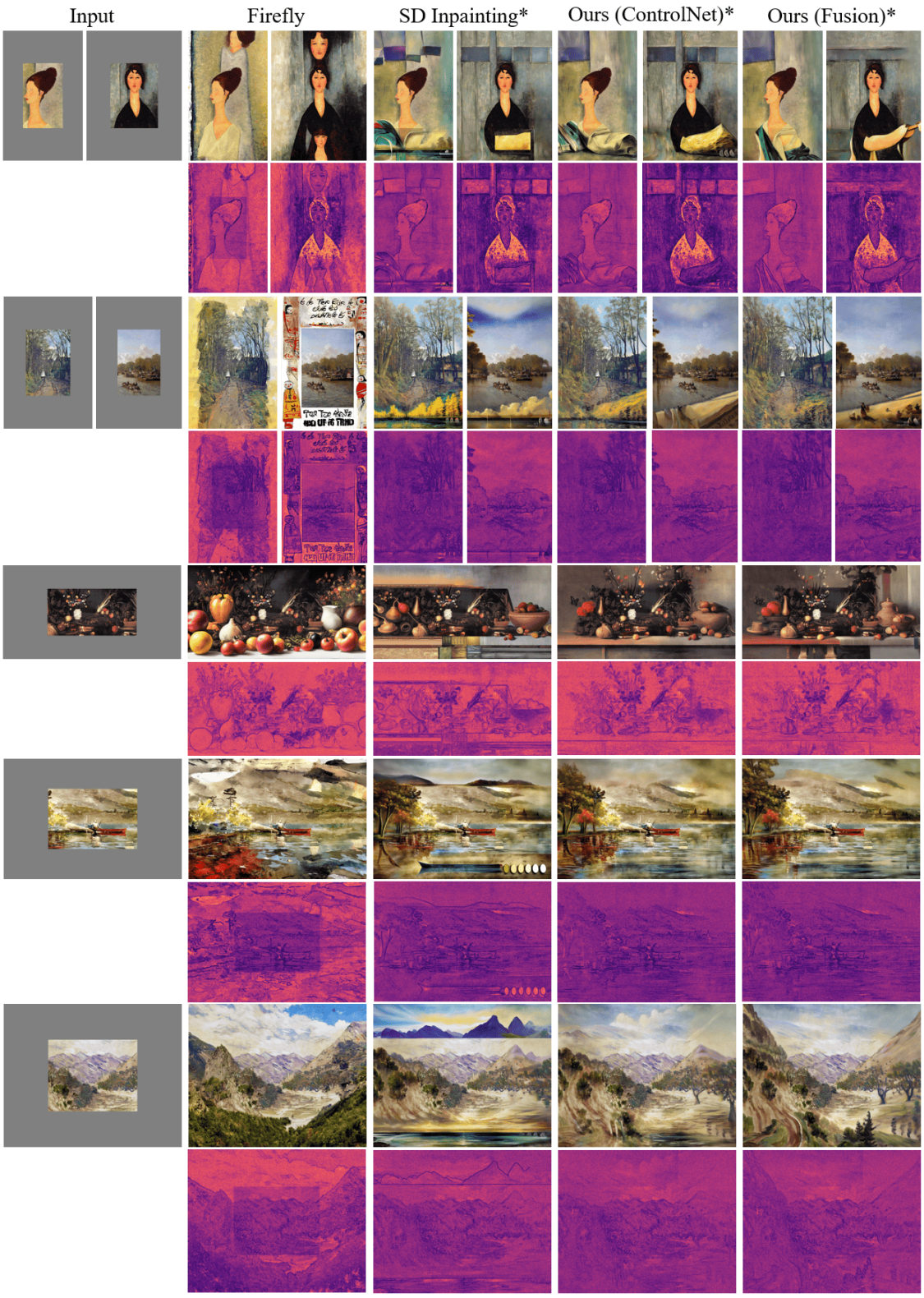

Figure 6. Blurriness visualization of the Figure 9 in the main paper. The darker the area of the map, the sharper it is, while the brighter the area, the blurriness becomes worse. An "*" indicates that the method based on the progressive outpainting, whereas Firefly generates in all directions at once.

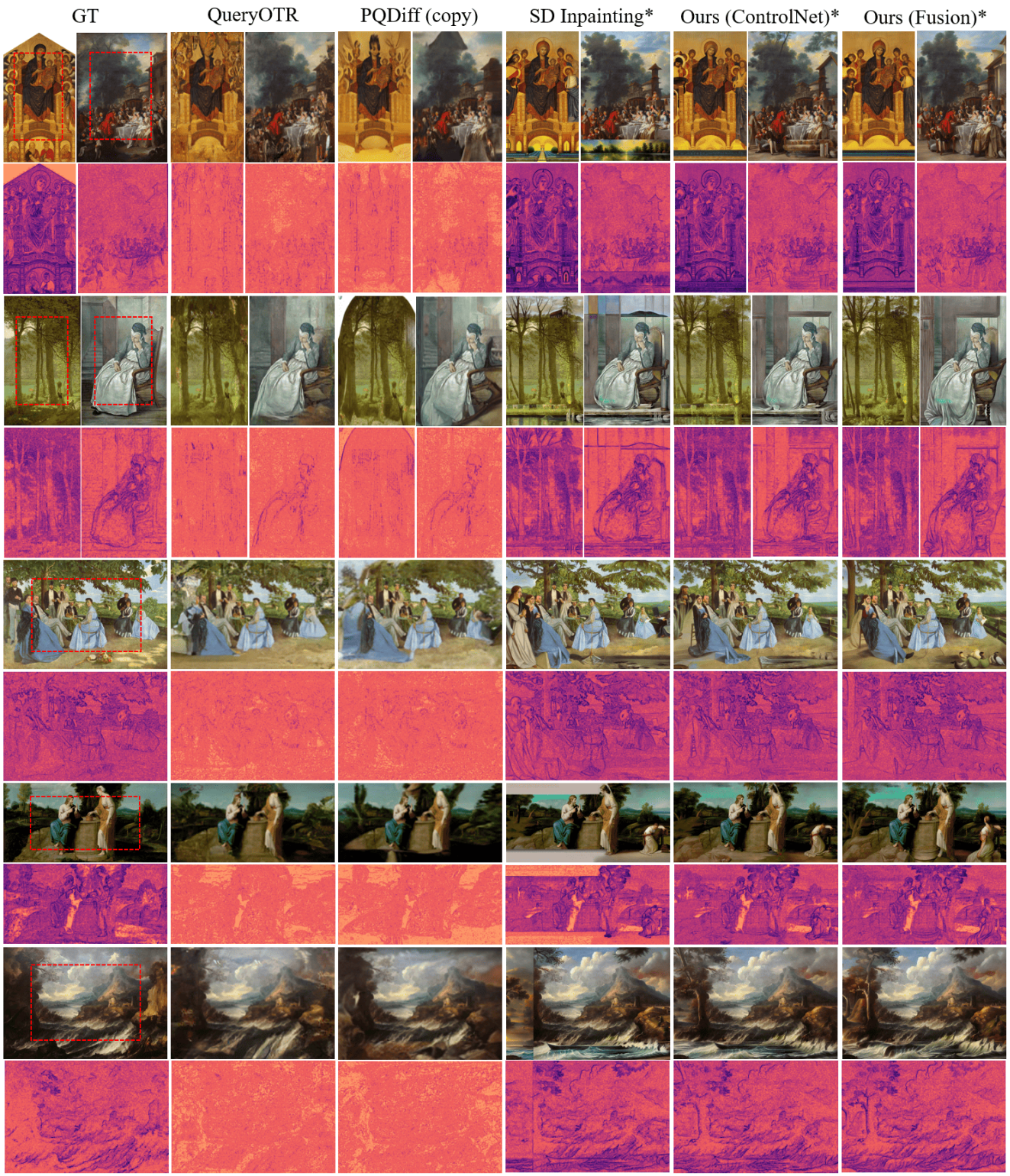

Figure 7. Qualitative results based on the scenario of Figure 7-(a) in the main paper. We report the quantitative results in the Table 1 of the main paper, using IconArt [4] dataset. The inner area of red boxes in GT denotes input pixels. The darker the area of the map, the sharper it is, while the brighter the area, the blurriness becomes worse. An "*" indicates that the method based on the progressive outpainting, whereas the others generate in all directions at once and are supported with super-resoultion.

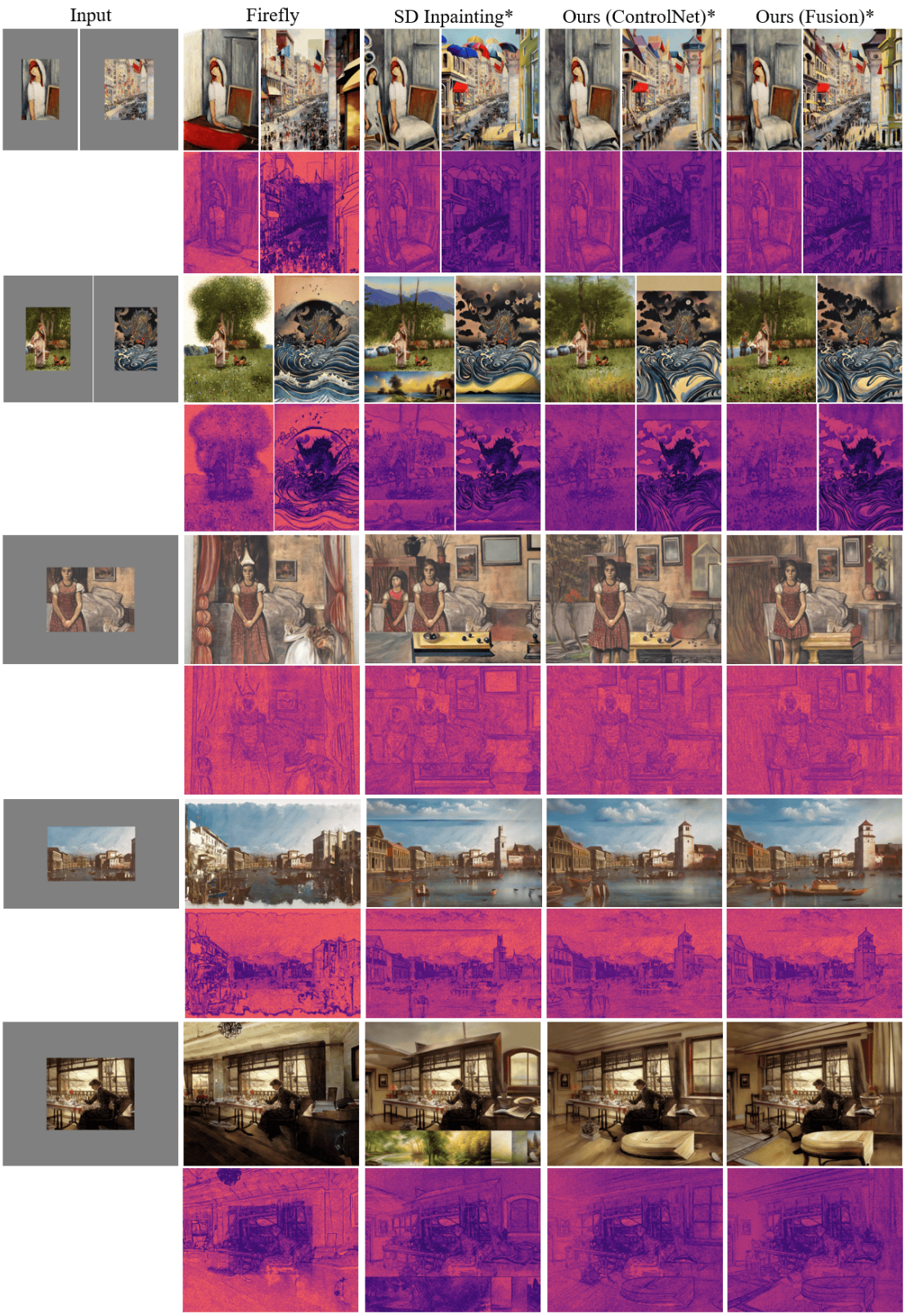

Figure 8. Qualitative results with corresponding blurriness maps based on the scenario of Figure 7-(b) in the main paper. We report the quantitative results in the Table 2 of the main paper, using WikiArt [12] dataset. The darker the area of the map, the sharper it is, while the brighter the area, the blurriness becomes worse. An "*" indicates that the method based on the progressive outpainting, whereas Firefly generates in all directions at once.

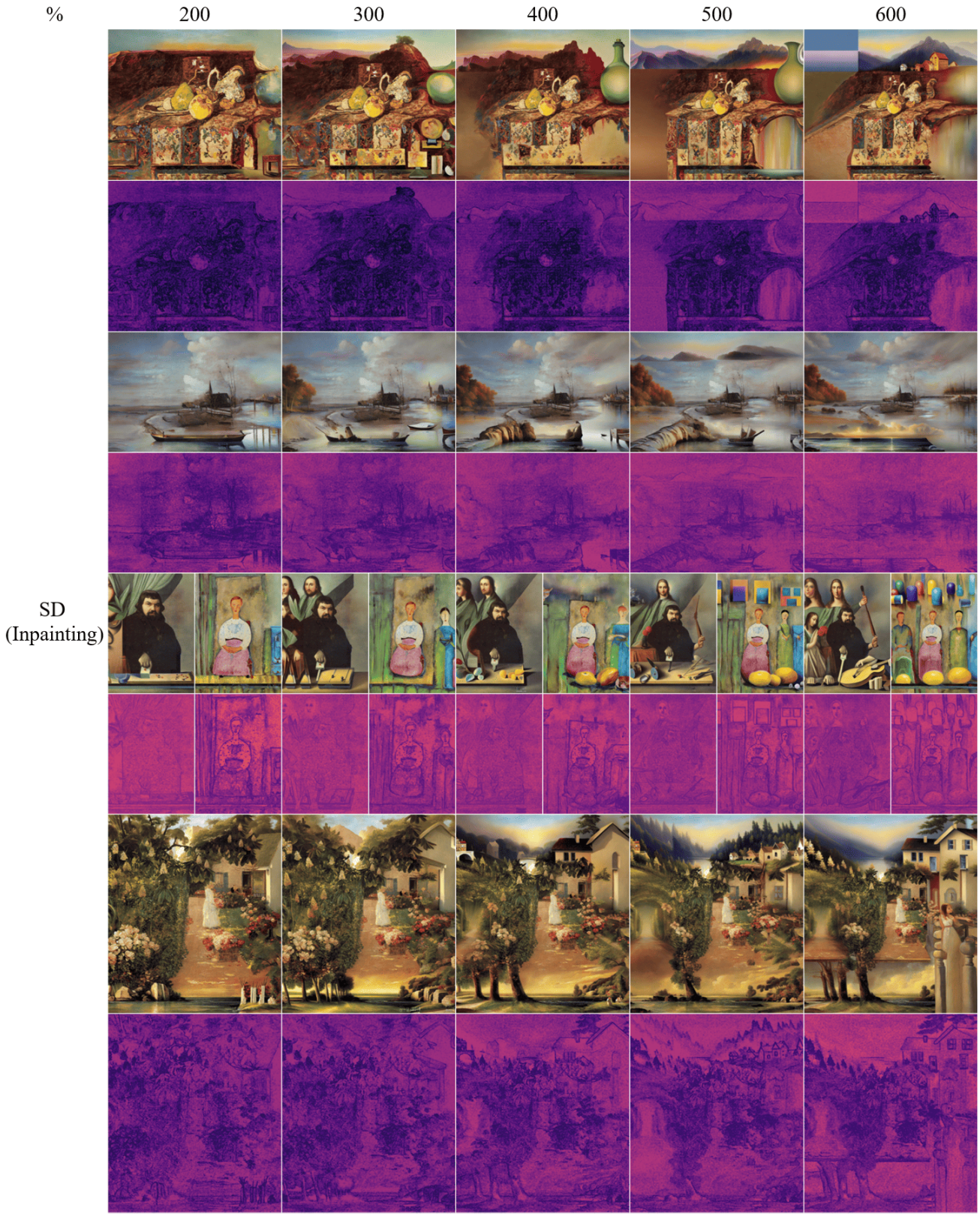

Figure 9. Qualitative outpainting results and blurriness visualization maps of the SD Inpainting model for the experiment corresponding to Table 3 in the main paper. The darker the area of the map, the sharper it is, while the brighter the area, the blurriness becomes worse.

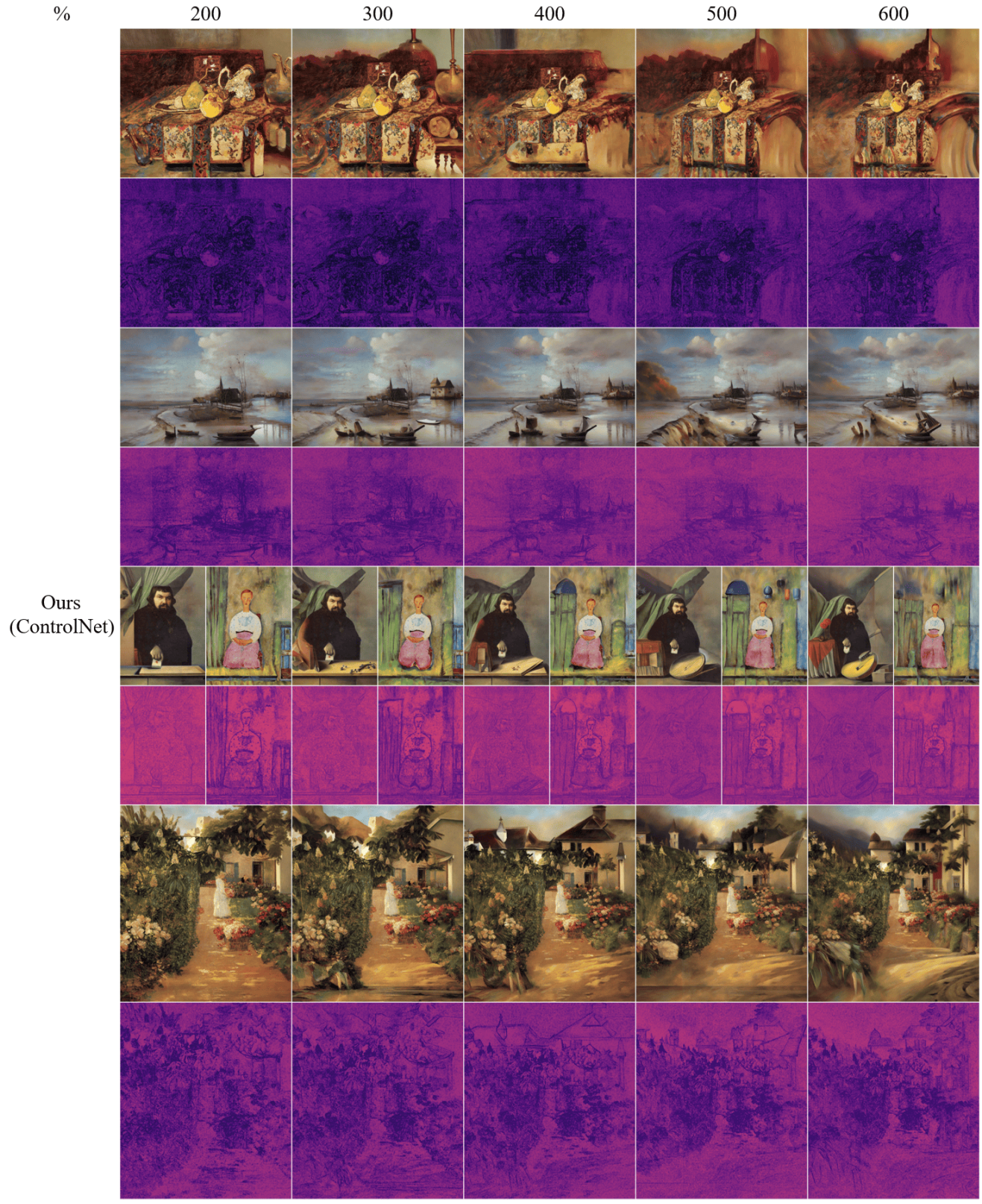

Figure 10. Qualitative outpainting results and blurriness visualization maps of the ControlNet-based CPM method for the experiment corresponding to Table 3 in the main paper. The darker the area of the map, the sharper it is, while the brighter the area, the blurriness becomes worse.

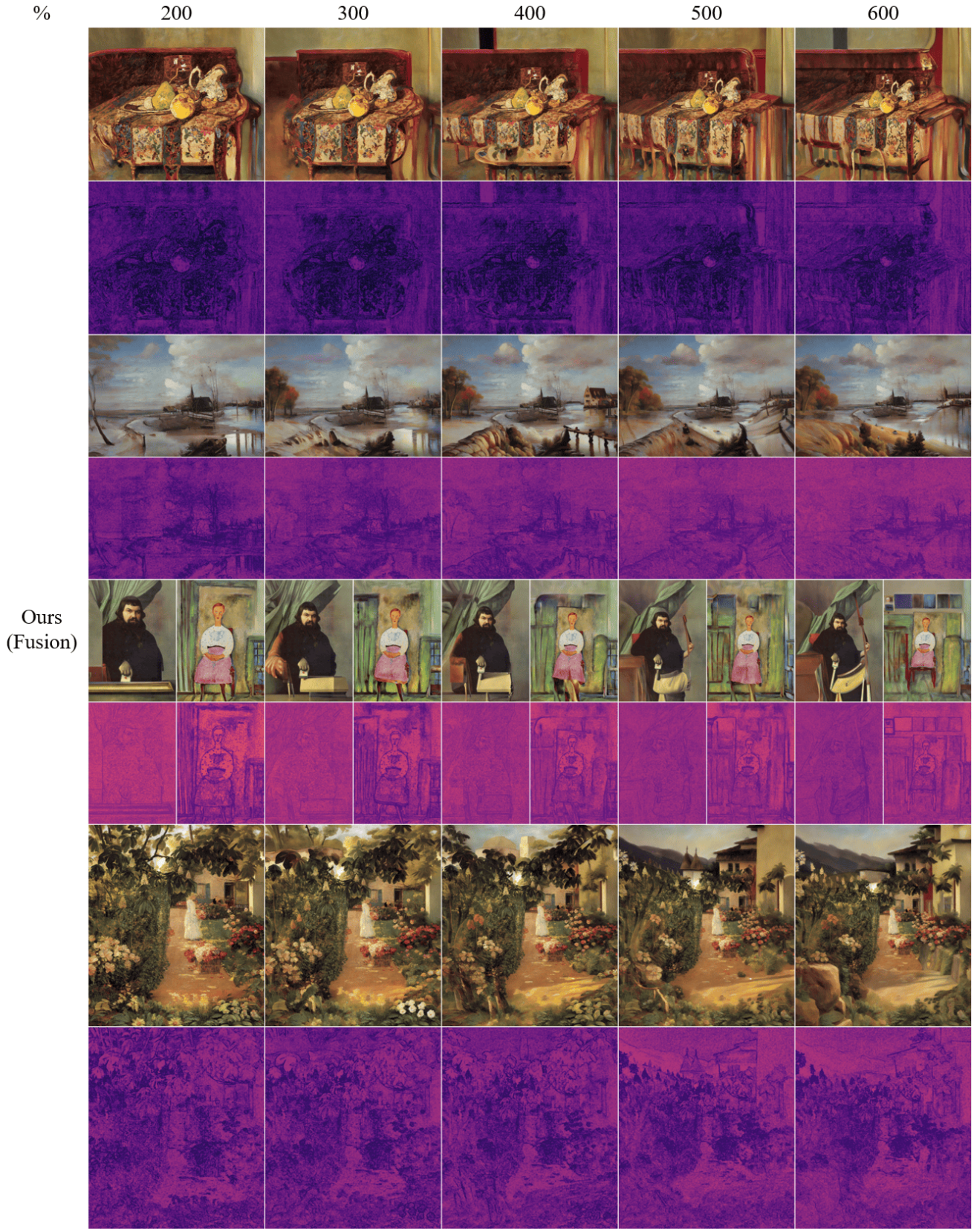

Figure 11. Qualitative outpainting results and blurriness visualization maps of the fusion-based CPM method for the experiment corresponding to Table 3 in the main paper. The darker the area of the map, the sharper it is, while the brighter the area, the blurriness becomes worse.

References

[1] Binxin Yang et al. Paint by example: Exemplar-based image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[2] Ivan Skorokhodov et al. Aligning latent and image spaces to connect the unconnectable. In *International Conference on Computer Vision (ICCV)*, 2021. 5

[3] Luming Tang et al. Realfill: Reference-driven generation for authentic image completion. *ACM Transactions on Graphics (ToG)*, 2024. 6

[4] Nicolas Gonthier, Yann Gousseau, Said Ladjal, and Olivier Bonfait. Weakly supervised object detection in artworks. In *European Conference on Computer Vision Workshop (ECCVW)*, pages 0–0, 2018. 9

[5] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017. 7

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, 2022. 5

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022. 3

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 7

[11] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems (NIPS) Datasets and Benchmarks Track*, 2022. 7

[12] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing (TIP)*, 28(1):394–409, 2019. 10

[13] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 4