CLAIMS:

1. A method comprising:

obtaining a user edit input for a source image, the user edit input at least indicating movement of a feature at at least one handle point to at least one target point in the source image;

extracting a feature map from the source image using a diffusion model, the feature map being extracted by the diffusion model at a iteration step of an inverse denoising diffusion process performed on the source image;

updating the feature map based on the user edit input, to obtain an updated feature map; and

generating a target image based on the updated feature map using the diffusion model, the target image being generated by the diffusion model through a denoising diffusion process performed on the updated feature map.

2. The method of claim 1, wherein the diffusion model comprises a base model and a low rank adaptation (LoRA) part connected to the base model, the method further comprising:

performing, based on the source image, a fine-tuning process on a parameter set of the LoRA part of the diffusion model, with a parameter set of the base model unchanged, to obtain an updated diffusion model, and

wherein the extraction of the feature map and the generation of the target image are based on the updated diffusion model.

3. The method of claim 2, wherein performing the fine-tuning process comprises:

performing the fine-tuning process based on a training objective, the training objective being configured to cause the updated diffusion model to reconstruct the source image from a noised image, the noised image being generated by adding a noise signal on the source image.

4. The method of claim 2, wherein the LoRA part comprises at least a part of one or more attention modules comprised in the diffusion model.

5. The method of claim 1, wherein updating the feature map based on the user edit input comprises:

updating the feature map in an iteration process based on an optimization objective, the optimization objective being configured to cause a difference between a first feature segment of a first patch around the at least one handle point in the feature map and a second feature segment of a second patch around the at least one target point in the updated feature map to be minimized or to be reduced to a target error.

6. The method of claim 5, wherein the user edit input further indicates an editable region in the source image, and wherein the optimization objective is configured to cause a feature segment corresponding to an region outside the editable region to be unchanged.

7. The method of claim 5, wherein updating the feature map in an iteration process comprises:

at each iteration step of the iteration process, determining a motion supervision loss by:

obtaining an intermediate feature map that is the feature map at an initial iteration step or an intermediate feature map updated from a previous iteration step,

extracting a first intermediate feature segment from a first patch of the intermediate feature map around the at least one handle point in the intermediate feature map;

extracting a second intermediate feature segment from a second patch of the intermediate feature map around at least one moved handle point in the intermediate feature map, the at least one moved handle point being moved from the at least one handle point towards the at least one target point by a predetermined distance;

determining a motion supervision loss based on a difference between the first intermediate feature segment and the second intermediate feature segment;

updating the intermediate feature map to generate an updated intermediate feature map for a next iteration step by performing gradient backpropagation using the motion supervision loss, wherein a gradient is not backpropagated through the first intermediate feature segment.

8. The method of claim 7, wherein updating the feature map in an iteration process comprises:

updating the at least one handle point for the updated intermediate feature map to be at least one point that is selected a patch among the at least one handle point in the feature map through a nearest neighbor search.

9. The method of claim 1, wherein the diffusion model comprises a self-attention module with a key, a value, and a query as inputs, and wherein generating a target image based on the updated feature map using the diffusion model comprises:

performing the denoising diffusion process on the feature map by the diffusion model, to obtain a first key, a first value, and a first query for the self-attention module;

performing the denoising diffusion process on the updated feature map by the diffusion model, to obtain a second key, a second value, and a second query for the self-attention module;

providing the first key, the first value, and the second query as inputs to the self-attention module;

generating a target image based on an output of the self-attention module.

10. The method of claim 1, wherein the inverse denoising diffusion process comprises a total number of iteration steps, and the feature map is extracted at an iteration step that is within a range of sixty percentage or eighty percentage of the total number of iteration steps.

11. An electronic device, comprising:

at least one processor; and

at least one memory communicatively coupled to the at least one processor and comprising computer-readable instructions that upon execution by the at least one processor cause the at least one processor to perform acts comprising:

obtaining a user edit input for a source image, the user edit input at least indicating movement of a feature at at least one handle point to at least one target point in the source image;

extracting a feature map from the source image using a diffusion model, the feature map being extracted by the diffusion model at a iteration step of an inverse denoising diffusion process performed on the source image;

updating the feature map based on the user edit input, to obtain an updated feature map; and

generating a target image based on the updated feature map using the diffusion model, the target image being generated by the diffusion model through a denoising diffusion process performed on the updated feature map.

12. The device of claim 11, wherein the diffusion model comprises a base model and a low rank adaptation (LoRA) part connected to the base model, the device further comprising:

performing, based on the source image, a fine-tuning process on a parameter set of the LoRA part of the diffusion model, with a parameter set of the base model unchanged, to obtain an updated diffusion model, and

wherein the extraction of the feature map and the generation of the target image are based on the updated diffusion model.

13. The device of claim 12, wherein performing the fine-tuning process comprises:

performing the fine-tuning process based on a training objective, the training objective being configured to cause the updated diffusion model to reconstruct the source image from a noised image, the noised image being generated by adding a noise signal on the source image.

14. The device of claim 12, wherein the LoRA part comprises at least a part of one or more attention modules comprised in the diffusion model.

15. The device of claim 11, wherein updating the feature map based on the user edit input comprises:

updating the feature map in an iteration process based on an optimization objective, the optimization objective being configured to cause a difference between a first feature segment of a first patch around the at least one handle point in the feature map and a second

feature segment of a second patch around the at least one target point in the updated feature map to be minimized or to be reduced to a target error.

16. The device of claim 15, wherein the user edit input further indicates an editable region in the source image, and wherein the optimization objective is configured to cause a feature segment corresponding to an region outside the editable region to be unchanged.

17. The device of claim 15, wherein updating the feature map in an iteration process comprises:

at each iteration step of the iteration process, determining a motion supervision loss by:

obtaining an intermediate feature map that is the feature map at an initial iteration step or an intermediate feature map updated from a previous iteration step,

extracting a first intermediate feature segment from a first patch of the intermediate feature map around the at least one handle point in the intermediate feature map;

extracting a second intermediate feature segment from a second patch of the intermediate feature map around at least one moved handle point in the intermediate feature map, the at least one moved handle point being moved from the at least one handle point towards the at least one target point by a predetermined distance;

determining a motion supervision loss based on a difference between the first intermediate feature segment and the second intermediate feature segment;

updating the intermediate feature map to generate an updated intermediate feature map for a next iteration step by performing gradient backpropagation using the motion supervision loss, wherein a gradient is not backpropagated through the first intermediate feature segment.

18. The device of claim 17, wherein updating the feature map in an iteration process comprises:

updating the at least one handle point for the updated intermediate feature map to be at least one point that is selected a patch among the at least one handle point in the feature map through a nearest neighbor search.

19. The device of claim 11, wherein the diffusion model comprises a self-attention module with a key, a value, and a query as inputs, and wherein generating a target image based on the updated feature map using the diffusion model comprises:

performing the denoising diffusion process on the feature map by the diffusion model, to obtain a first key, a first value, and a first query for the self-attention module;

performing the denoising diffusion process on the updated feature map by the diffusion model, to obtain a second key, a second value, and a second query for the self-attention module;

providing the first key, the first value, and the second query as inputs to the self-attention module;

generating a target image based on an output of the self-attention module.

20. A non-transitory computer-readable storage medium, storing computer-readable instructions that upon execution by a device cause the device to perform acts comprising:

obtaining a user edit input for a source image, the user edit input at least indicating movement of a feature at at least one handle point to at least one target point in the source image;

extracting a feature map from the source image using a diffusion model, the feature map being extracted by the diffusion model at a iteration step of an inverse denoising diffusion process performed on the source image;

updating the feature map based on the user edit input, to obtain an updated feature map; and

generating a target image based on the updated feature map using the diffusion model, the target image being generated by the diffusion model through a denoising diffusion process performed on the updated feature map.