WHAT IS CLAIMED IS:

1. A method for generating a video, comprising:

acquiring face images and an audio clip corresponding to each face image of the face images;

extracting face shape information and head posture information from the each face image, acquiring facial expression information according to the audio clip corresponding to the each face image, and acquiring face key point information of the each face image according to the facial expression information, the face shape information, and the head posture information;

inpainting, according to the face key point information of the each face image, the face images acquired, acquiring each generated image; and

generating a target video according to the each generated image.

2. The method of claim 1, wherein acquiring the face key point information of the each face image according to the facial expression information, the face shape information, and the head posture information comprises:

acquiring face point cloud data according to the facial expression information and the face shape information; and projecting the face point cloud data to a two-dimensional image according to the head posture information to obtain the face key point information of the each face image.

3. The method of claim 1 or 2, wherein acquiring the facial expression information according to the audio clip corresponding to the each face image comprises:

extracting an audio feature of the audio clip; removing timbre information of the audio feature; and acquiring the facial expression information according to the audio feature with the timbre information removed.

4. The method of claim 3, wherein removing the timbre information of the audio feature comprises:

removing the timbre information of the audio feature by normalizing the audio feature.

5. The method of claim 1 or 2, wherein generating the target video according to the each generated image comprises:

adjusting, according to the face images acquired, a regional image of the each generated image other than a face key point to obtain an adjusted generated image, and forming the target video using the adjusted generated image.

6. The method of claim 1 or 2, further comprising:

performing motion smoothing processing on a face key point of a speech-related part of an image in the target video, and / or performing jitter elimination on the image in the target video, wherein the speech-related part comprises at least a mouth and a chin.

7. The method of claim 6, wherein performing motion smoothing processing on the face key point of the speech-related part of the image in the target video comprises:

for a t greater than or equal to 2, in response to a distance between a center of a speech-related part of a t-th image of the target video and a center of a speech-related part of a (t-1)-th image of the target video being less than or equal to a set distance threshold, acquiring motion smoothed face key point information of the speech-related part of the t-th image of the target video according to face key point information of the speech-related part of the t-th image of the target video and face key point information of the speech-related part of the (t-1)-th image of the target video.

8. The method of claim 6, wherein performing jitter elimination on the image in the target video comprises:

for a t greater than or equal to 2, performing jitter elimination on a t-th image of the target video according to a light flow from a (t-1)-th image of the target video to the t-th

image of the target video, the (t-1)-th image of the target video with jitter eliminated, and a distance between a center of a speech-related part of the t-th image of the target video and a center of a speech-related part of the (t-1)-th image of the target video.

9. The method of claim 1 or 2, wherein acquiring the face images and the audio clip corresponding to the each face image of the face images comprises:

acquiring source video data, separating the face images and audio data comprising a voice from the source video data, and determining the audio clip corresponding to the each face image, the audio clip corresponding to the each face image being part of the audio data.

10. The method of claim 1 or 2, wherein extracting the face shape information and the head posture information from the each face image, acquiring the facial expression information according to the audio clip corresponding to the each face image, and acquiring the face key point information of the each face image according to the facial expression information, the face shape information, and the head posture information comprises:

inputting the face images and the audio clip corresponding to the each face image to a first neural network trained in advance; and extracting the face shape information and the head posture information from the each face image, acquiring the facial expression information according to the audio clip corresponding to the each face image, and acquiring the face key point information of the each face image according to the facial expression information, the face shape information, and the head posture information based on the first neural network.

11. The method of claim 10, wherein the first neural network is trained by:

acquiring multiple sample face images and a sample audio clip corresponding to each sample face image of the multiple sample face images;

inputting the each sample face image and the sample audio clip corresponding to the each sample face image to the first neural network yet to be trained, acquiring predicted facial expression information and predicted face key point information of the each sample face image;

adjusting a network parameter of the first neural network according to a loss of the first neural network, the loss of the first neural network comprising an expression loss and / or a face key point loss, the expression loss being configured to represent a difference between the predicted facial expression information and a facial expression marker result, the face key point loss being configured to represent a difference between the predicted face key point information and a face key point marker result; and

repeating above-mentioned steps until the loss of the first neural network meets a first predetermined condition, acquiring the first neural network that has been trained.

12. The method of claim 1 or 2, wherein inpainting, according to the face key point information of the each face image, the face images acquired, acquiring the each generated image comprises:

inputting the face key point information of the each face image and the face images acquired to a second neural network trained in advance, and inpainting, based on the second neural network according to the face key point information of the each face image, the face images acquired, to obtain the each generated image.

13. The method of claim 12, wherein the second neural network is trained by:

acquiring a face image with a masked portion by adding a mask to a sample face image with no masked portion acquired in advance, inputting sample face key point information acquired in advance and the face image with the masked portion to the second neural network yet to be trained, and inpainting, according to the sample face key point information based on the second neural network, the masked portion of the face image with the masked portion, to obtain a generated image;

discriminating the sample face image to obtain a first discrimination result, and discriminating the generated image to obtain a second discrimination result;

adjusting a network parameter of the second neural network according to a loss of the second neural network, the loss of the second neural network comprising an adversarial loss, the adversarial loss being acquired according to the first discrimination result and the second discrimination result; and

repeating above-mentioned steps until the loss of the second neural network meets a second predetermined condition, acquiring the second neural network that has been trained.

14. The method of claim 13, wherein the loss of the second neural network further comprises at least one of a pixel reconstruction loss, a perceptual loss, an artifact loss, or a gradient penalty loss, the pixel reconstruction loss being configured to represent a difference between the sample face image and the generated image, the perceptual loss being configured to represent a sum of differences between the sample face image and the generated image at different scales, the artifact loss being configured to represent a spike artifact of the generated image, the gradient penalty loss being configured to limit a gradient for updating the second neural network.

15. A device for generating a video, comprising a first processing module, a second processing module, and a generating module,

wherein the first processing module is configured to acquire face images and an audio clip corresponding to each face image of the face images,

wherein the second processing module is configured to extract face shape information and head posture information from the each face image, acquire facial expression information according to the audio clip corresponding to the each face image, and acquire face key point information of the each face image according to the facial expression information, the face shape information, and the head posture information; inpaint, according to the face key point information of the each face image, the face images acquired, acquiring each generated image,

wherein the generating module is configured to generate a target video according to the each generated image.

16. The device of claim 15, wherein the second processing module is configured to acquire face point cloud data according to the facial expression information and the face shape information; and project the face point cloud data to a two-dimensional image according to the head posture information to obtain the face key point information of the each face image.

17. The device of claim 15 or 16, wherein the second processing module is configured to extract an audio feature of the audio clip; remove timbre information of the audio feature; and acquire the facial expression information according to the audio feature with the timbre information removed.

~~18. The device of claim 17, wherein the second processing module is configured to remove the timbre information of the audio feature by normalizing the audio feature.~~

~~19. The device of claim 15 or 16, wherein the generating module is configured to adjust, according to the face images acquired, a regional image of the each generated image other than a face key point to obtain an adjusted generated image, and form the target video using the adjusted generated image.~~

~~20. The device of claim 15 or 16, further comprising a jitter eliminating module,~~

~~wherein the jitter eliminating module is configured to perform motion smoothing processing on a face key point of a speech-related part of an image in the target video, and / or perform jitter elimination on the image in the target video, wherein the speech-related part comprises at least a mouth and a chin.~~

~~21. The device of claim 20, wherein the jitter eliminating module is configured to, for a t greater than or equal to 2, in response to a distance between a center of a speech-related part of a t-th image of the target video and a center of a speech-related part of a (t-1)-th image of the target video being less than or equal to a set distance threshold, acquire motion smoothed face key point information of the speech-related part of the t-th image of the target video according to face key point information of the speech-related part of the t th image of the target video and face key point information of the speech-related part of the (t-1)-th image of the target video.~~

~~22. The device of claim 20, wherein the jitter eliminating module is configured to, for a t~~

~~greater than or equal to 2, perform jitter elimination on a t-th image of the target video according to a light flow from a (t-1)-th image of the target video to the t-th image of the target video, the (t-1)-th image of the target video with jitter eliminated, and a distance between a center of a speech-related part of the t-th image of the target video and a center of a speech-related part of the (t-1)-th image of the target video.~~

~~23. The device of claim 15 or 16, wherein the first processing module is configured to acquire source video data, separate the face images and audio data comprising a voice from the source video data, and determine the audio clip corresponding to the each face image, the audio clip corresponding to the each face image being part of the audio data.~~

~~24. The device of claim 15 or 16, wherein the second processing module is configured to input the face images and the audio clip corresponding to the each face image to a first neural network trained in advance; and extract the face shape information and the head posture information from the each face image, acquire the facial expression information according to the audio clip corresponding to the each face image, and acquire the face key point information of the each face image according to the facial expression information, the face shape information, and the head posture information based on the first neural network.~~

~~25. The device of claim 24, wherein the first neural network is trained by:~~

~~acquiring multiple sample face images and a sample audio clip corresponding to each sample face image of the multiple sample face images;~~

~~inputting the each sample face image and the sample audio clip corresponding to the each sample face image to the first neural network yet to be trained, acquiring predicted facial expression information and predicted face key point information of the each sample face image;~~

~~adjusting a network parameter of the first neural network according to a loss of the first neural network, the loss of the first neural network comprising an expression loss and / or a face key point loss, the expression loss being configured to represent a difference between the predicted facial expression information and a facial expression marker result, the~~

face key point loss being configured to represent a difference between the predicted face key point information and a face key point marker result; and

repeating above-mentioned steps until the loss of the first neural network meets a first predetermined condition, acquiring the first neural network that has been trained.

26. The device of claim 15 or 16, wherein the second processing module is configured to input the face key point information of the each face image and the face images acquired to a second neural network trained in advance, and inpaint, based on the second neural network according to the face key point information of the each face image, the face images acquired, to obtain the each generated image.

27. The device of claim 26, wherein the second neural network is trained by:

acquiring a face image with a masked portion by adding a mask to a sample face image with no masked portion acquired in advance, inputting sample face key point information acquired in advance and the face image with the masked portion to the second neural network yet to be trained, and inpainting, according to the sample face key point information based on the second neural network, the masked portion of the face image with the masked portion, to obtain a generated image;

discriminating the sample face image to obtain a first discrimination result, and discriminating the generated image to obtain a second discrimination result;

adjusting a network parameter of the second neural network according to a loss of the second neural network, the loss of the second neural network comprising an adversarial loss, the adversarial loss being acquired according to the first discrimination result and the second discrimination result; and

repeating above-mentioned steps until the loss of the second neural network meets a second predetermined condition, acquiring the second neural network that has been trained.

28. The device of claim 27, wherein the loss of the second neural network further comprises at least one of a pixel reconstruction loss, a perceptual loss, an artifact loss, or a gradient penalty loss, the pixel reconstruction loss being configured to represent a difference between

~~the sample face image and the generated image, the perceptual loss being configured to represent a sum of differences between the sample face image and the generated image at different scales, the artifact loss being configured to represent a spike artifact of the generated image, the gradient penalty loss being configured to limit a gradient for updating the second neural network.~~

18~~29~~. Electronic equipment, comprising a processor and a memory configured to store a computer program executable on the processor,

wherein the processor is configured to implement the method of any one of claims 1 to 14 when executing the computer program.

19~~30~~. A computer-readable storage medium, having stored thereon a computer program which, when executed by a processor, implements the method of any one of claims 1 to 14.

20~~31~~. A computer program, comprising a computer-readable code which, when executed in electronic equipment, allows a processor in the electronic equipment to implement the method of any one of claims 1 to 14.