CLAIMS

What is claimed is:

1. A computer-implemented neural network construction method, comprising:

obtaining a start point network, wherein the start point network comprises a plurality of serial subnets;

performing at least one time of transformation on the start point network based on a first search space to obtain a serial network, wherein the first search space comprises a range of parameters used for transforming the start point network;

if the serial network meets a preset condition, training the serial network by using a preset dataset to obtain a trained serial network; and

if the trained serial network meets a termination condition, obtaining a target neural network based on the trained serial network and using the target neural network to perform at least one of image recognition, semantic segmentation, or object detection, wherein the preset dataset comprises a dataset comprising image data;

wherein each serial subnet in the start point network comprises one or more blocks, and the transforming the start point network comprises:

swapping blocks comprised in two adjacent serial subnets in the plurality of serial subnets, the adjacent serial subnets divided by downsampling, the swapping comprising swapping a location of the downsampling and a location of a block in one of the two adjacent serial subnets; or

inserting a processed block into at least one serial subnet in the plurality of serial subnets, wherein the processed block is obtained by initializing a weight of a block as an identity matrix and performing group normalization on the block.

2. The computer-implemented method according to claim 1, wherein one of the at least one time of transformation performed on the start point network based on the first search space comprises:

searching the first search space for a transformation policy, wherein the transformation policy comprises a group of parameters used for transforming the start point network;

transforming the start point network according to the transformation policy to obtain a transformed start point network, wherein a weight of the transformed start point network inherits

a weight of the start point network before transformation;

updating the weight of the transformed start point network by using a lightweight dataset, to obtain a start point network with an updated weight, wherein data comprised in the preset dataset is more than data comprised in the lightweight dataset;

evaluating the start point network with the updated weight to obtain a first evaluation result; and

if the first evaluation result meets a preset update condition, re-searching the first search space for a transformation policy, and re-transforming the start point network according to the re-found transformation policy; or

if the first evaluation result does not meet the preset update condition, using the start point network with the updated weight as the serial network.

3. The computer-implemented method according to claim 2, wherein the method further comprises:

if the serial network does not meet the preset condition, using the serial network as a new start point network; or if the trained serial network does not meet the termination condition, using the trained serial network as a new start point network; and

updating the first search space to include a range of parameters used for transforming the new start point network, searching for a transformation policy based on an updated first search space, and transforming the new start point network according to the found transformation policy.

4. The computer-implemented method according to claim 2 or 3, wherein the preset update condition comprises: a quantity of times of transformation performed on the start point network does not exceed a preset quantity of times, or the first evaluation result does not meet a preset convergence condition, wherein the preset convergence condition comprises one or more of the following: the first evaluation result is better than a threshold, or a difference between the first evaluation result and an evaluation result obtained in previous evaluation of the start point network is less than a preset value.

5. The computer-implemented method according to any one of claims 2 to 4, wherein the preset condition comprises: a second evaluation result of the serial network meets a first preset evaluation condition, a quantity of times of transformation performed on the serial network exceeds a preset quantity of times of transformation, or duration from last time of transformation performed on the start point network exceeds a preset value, wherein the second evaluation result

is obtained by evaluating the serial network; and

the first preset evaluation condition comprises one or more of the following: a loss value of the serial network is not greater than a first loss threshold, a precision value of an output result of the serial network is greater than a first precision threshold, or inference duration of the serial network is less than a first inference threshold.

6. The computer-implemented method according to claim 5, wherein the termination condition comprises:

the second evaluation result of the serial network meets a second preset evaluation condition, or a quantity of times of training of the serial network exceeds a preset quantity of times of training.

7. The computer-implemented method according to any one of claims 1 to 6, wherein the obtaining a target neural network based on the trained serial network comprises:

if the trained serial network meets an output condition, using the trained serial network as the target neural network; or

if the trained serial network does not meet the output condition, obtaining a parallel subnet corresponding to the trained serial network, and constructing the target neural network based on the trained serial network and the parallel subnet corresponding to the trained serial network.

8. The computer-implemented method according to claim 7, wherein the serial network comprises a plurality of serial subnets, and each serial subnet comprises one or more blocks; and

the obtaining a parallel subnet corresponding to the trained serial network comprises:

determining, within a preset quantity range, a quantity of parallel subnets corresponding to each serial subnet in the plurality of serial subnets; and

constructing, based on the quantity of parallel subnets corresponding to each serial subnet, the parallel subnet corresponding to each serial subnet, wherein a structure of each serial subnet is the same as a structure of the parallel subnet corresponding to each serial subnet.

9. The computer-implemented method according to claim 8, wherein the constructing the target neural network based on the trained serial network and the parallel subnet corresponding to the trained serial network comprises:

constructing a temporary neural network based on the trained serial network and the parallel subnet corresponding to each serial subnet;

evaluating the temporary neural network to obtain a third evaluation result; and

if the third evaluation result meets the output condition, using the temporary neural network

as the target neural network; or

if the third evaluation result does not meet the output condition, re-obtaining a parallel subnet corresponding to each serial subnet.

10. The computer-implemented method according to any one of claims 7 to 9, wherein the output condition comprises one or more of the following: precision is greater than a first threshold; average precision is greater than a second threshold; the loss value is not greater than a third threshold; or the inference duration is not greater than a fourth threshold, and floating-point operations per second is not greater than a fifth threshold, wherein the average precision is an average value of a plurality of precisions obtained by evaluating a neural network for a plurality of times, and the inference duration is duration for obtaining an output result from the neural network based on an input.

11. The computer-implemented method according to any one of claims 1 to 10, wherein types of parameters comprised in the first search space comprise: a width, a depth, a quantity of times of downsampling, and a location of downsampling, wherein the width comprises a quantity of input channels and a quantity of output channels of a block comprised in each serial subnet in the start point network, the depth is a quantity of blocks comprised in each serial subnet in the start point network, and downsampling is used to divide two adjacent serial subnets in the start point network.

12. The method according to any one of claims 1 to 12, wherein the preset dataset comprises an ImageNet dataset.

13. A neural network construction apparatus, comprising:

an obtaining module, configured to obtain a start point network, wherein the start point network comprises a plurality of serial subnets;

a transformation module, configured to perform at least one time of transformation on the start point network based on a first search space to obtain a serial network, wherein the first search space comprises a range of parameters used for transforming the start point network;

a training module, configured to: if the serial network meets a preset condition, train the serial network by using a preset dataset to obtain a trained serial network; and

a construction module, configured to: if the trained serial network meets a termination condition, obtain a target neural network based on the trained serial network; and

a processor configured to use the target neural network to perform at least one of image

recognition, semantic segmentation, or object detection, wherein the preset dataset comprises a dataset comprising image data;

wherein each serial subnet in the start point network comprises one or more blocks, and the transformation module is further configured to transform the start point network by:

swapping blocks comprised in two adjacent serial subnets in the plurality of serial subnets, the adjacent serial subnets divided by downsampling, the swapping comprising swapping a location of the downsampling and a location of a block in one of the two adjacent serial subnets; or

inserting a processed block into at least one serial subnet in the plurality of serial subnets, wherein the processed block is obtained by initializing a weight of a block as an identity matrix and performing group normalization on the block.

14. The neural network construction apparatus according to claim 13, wherein the transformation module is specifically configured to:

search the first search space for a transformation policy, wherein the transformation policy comprises a group of parameters used for transforming the start point network;

transform the start point network according to the transformation policy to obtain a transformed start point network, wherein a weight of the transformed start point network inherits a weight of the start point network before transformation;

update the weight of the transformed start point network by using a lightweight dataset, to obtain a start point network with an updated weight, wherein data comprised in the preset dataset is more than data comprised in the lightweight dataset;

evaluate the start point network with the updated weight to obtain a first evaluation result; and

if the first evaluation result meets a preset update condition, re-search the first search space for a transformation policy, and re-transform the start point network according to the re-found transformation policy; or

if the first evaluation result does not meet the preset update condition, use the start point network with the updated weight as the serial network.

15. The neural network construction apparatus according to claim 14, wherein the neural network construction apparatus further comprises:

an update module, configured to: if the serial network does not meet the preset condition, use

the serial network as a new start point network; or if the trained serial network does not meet the termination condition, use the trained serial network as a new start point network; and update the first search space to include a range of parameters used for transforming the new start point network; and

the transformation module is further configured to: search for a transformation policy based on an updated first search space, and transform the new start point network according to the found transformation policy.

16. The neural network construction apparatus according to claim 14 or 15, wherein the update condition comprises: a quantity of times of transformation performed on the start point network does not exceed a preset quantity of times, or the first evaluation result does not meet a preset convergence condition, wherein the convergence condition comprises one or more of the following: the first evaluation result is better than a threshold, or a difference between the first evaluation result and an evaluation result obtained in previous evaluation of the start point network is less than a preset value.

17. The neural network construction apparatus according to any one of claims 14 to 16, wherein the preset condition comprises:

a second evaluation result of the serial network meets a first preset evaluation condition, a quantity of times of transformation performed on the serial network exceeds a preset quantity of times of transformation, or duration from last time of transformation performed on the start point network exceeds a preset value, wherein the second evaluation result is obtained by evaluating the serial network; and

the first preset evaluation condition comprises one or more of the following: a loss value of the serial network is not greater than a first loss threshold, a precision value of an output result of the serial network is greater than a first precision threshold, or inference duration of the serial network is less than a first inference threshold.

18. A neural network construction apparatus, comprising a processor, wherein the processor is coupled to a memory, the memory stores a program instructions, and when the program instructions stored in the memory are executed by the processor, the method according to any one of claims 1 to 12 is implemented.

19. A computer-readable storage medium, comprising a program, wherein when the program is executed by a processing unit, the method according to any one of claims 1 to 12 is performed.

20. A neural network construction apparatus, comprising a processing unit and a communication interface, wherein the processing unit obtains program instructions through the communication interface, and when the program instructions are executed by the processing unit, the method according to any one of claims 1 to 12 is implemented.