



(51) International Patent Classification:

H04N 19/62 (2014.01)

(21) International Application Number:

PCT/US2020/060675

(22) International Filing Date:

16 November 2020 (16.11.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

16/684,412 14 November 2019 (14.11.2019) US

(71) Applicant: **ALIBABA GROUP HOLDING LIMITED**

[—/US]; Fourth Floor, One Capital Place, P.O. Box 847,
George Town, Grand Cayman (KY).

(72) Inventors: **XU, Kai**; Alibaba Group Legal Department,

5/F, Building 3, No. 969 West Wen Yi Road, Yu Hang

District, Hangzhou, 311121 (CN). **QIN, Minghai**; Alibaba Group Legal Department, 5/F, Building 3, No. 969 West Wen Yi Road, Yu Hang District, Hangzhou, 311121 (CN). **WANG, Yuhao**; Alibaba Group Legal Department, 5/F, Building 3, No. 969 West Wen Yi Road, Yu Hang District, Hangzhou, 311121 (CN). **SUN, Fei**; Alibaba Group Legal Department, 5/F, Building 3, No. 969 West Wen Yi Road, Yu Hang District, Hangzhou, 311121 (CN). **YEN-KUANG, Chen**; Alibaba Group Legal Department, 5/F, Building 3, No. 969 West Wen Yi Road, Yu Hang District, Hangzhou, 311121 (CN). **XIE, Yuan**; Alibaba Group Legal Department, 5/F, Building 3, No. 969 West Wen Yi Road, Yu Hang District, Hangzhou, 311121 (CN).

(74) Agent: **MURABITO, Anthony, C.**; Murabito Hao &

Barnes, LLP, 111 North Market St., #700, San Jose, CA 95113 (US).

(81) Designated States (*unless otherwise indicated, for every*

kind of national protection available): AE, AG, AL, AM,

(54) Title: TECHNIQUES TO DYNAMICALLY GATE ENCODED IMAGE COMPONENTS FOR ARTIFICIAL INTELLIGENCE TASKS

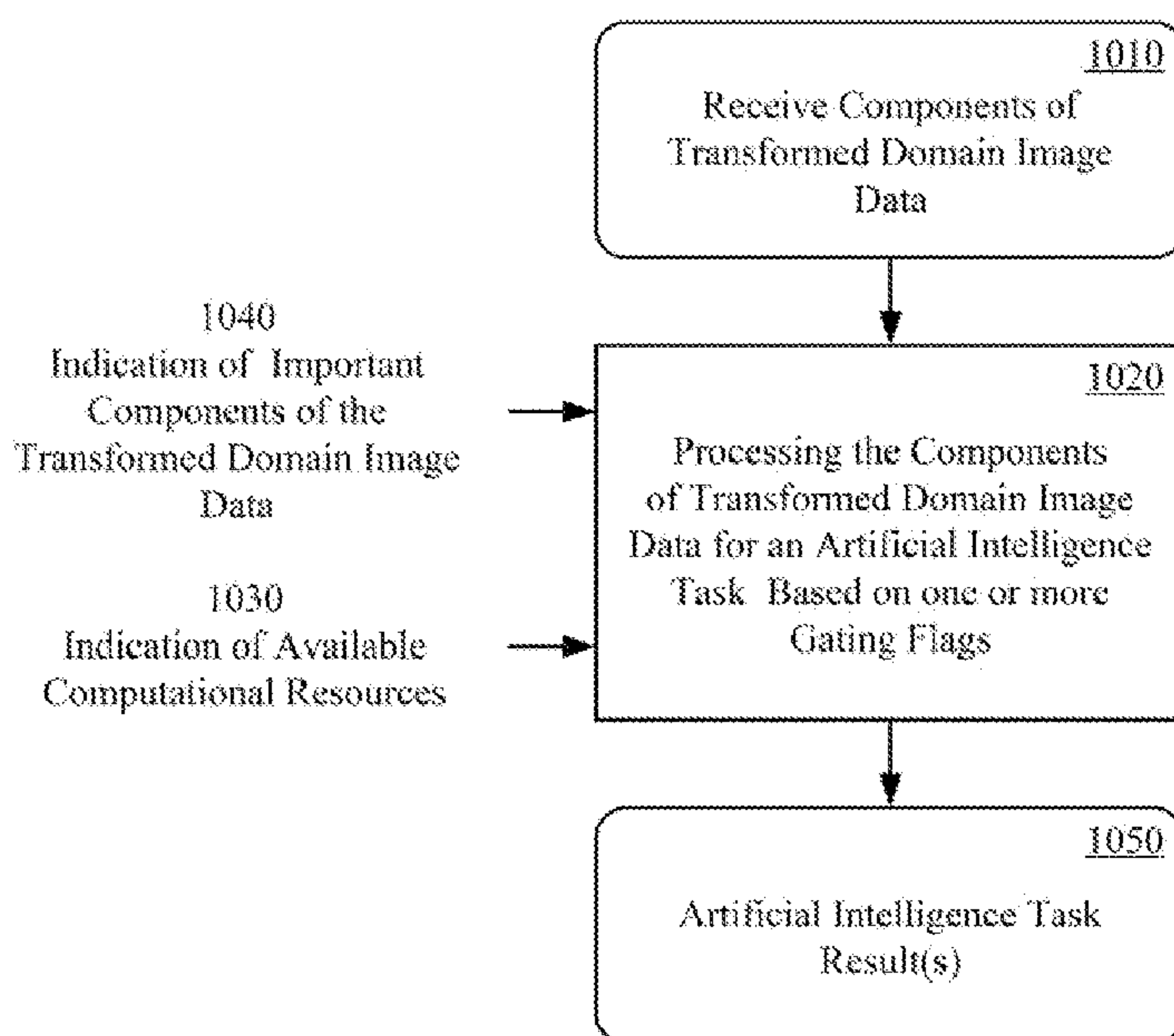


FIG. 10

(57) Abstract: A system for processing encoded image components for artificial intelligence tasks. The system can include one or more compute units, one or more controllers and memory. The one or more controllers can include one or more micro-op schedulers and one or more channel switches. The one or more compute units can be configured to process components of the transformed domain image data according to one or more micro-operations for an artificial intelligence task. The one or more channel switches can be configured to selectively control the transfer of the components of transformed domain image data to the one or more compute units based on one or more gating flags. The one or more channel switches can also be configured to selectively control generation of the one or more micro-operations by the one or more micro-op schedulers based on the one or more gating flags.

AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available):* ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

TECHNIQUES TO DYNAMICALLY GATE ENCODED IMAGE
COMPONENTS FOR ARTIFICIAL INTELLIGENCE TASKS

RELATED APPLICATIONS

5 **[0001]** This application is related to U.S. Patent Application No. _____, Filed November 14, 2019, Attorney Docket No. BABA-A24876 and entitled "Techniques for Determining Importance of Encoded Image Components for Artificial Intelligence Tasks," U.S. Patent Application No. _____, filed November 14, 2019, Attorney Docket No. BABA-A24878 and entitled "Reconstructing Transformed Domain Information in Encoded
10 Video Streams," and U.S. Patent Application No. _____, filed November 14, 2019, Attorney Docket No. BABA-A24879 and entitled "Using Selected Frequency Domain Components of Image Data in Artificial Intelligence Tasks," all of which are incorporated herein in their entirety.

BACKGROUND OF THE INVENTION

15 **[0002]** Artificial intelligence (AI), machine learning, and deep learning are utilized for various image processing tasks, computer vision tasks, and the like. Artificial intelligence as used herein refers to techniques that enable devices to mimic human intelligence, using logic, if-then rules, decision trees, and the like. Machine learning includes a subset of artificial intelligence that includes abstruse statistical techniques that enable machines to
20 improve at task with experience. Deep learning includes a subset of machine learning that includes algorithms that permit software to train itself to perform tasks by exposing multilayered artificial neural networks, recurrent neural networks (RNN), convolution neural networks (CNN) or the like to vast amounts of data. For ease of explanation artificial intelligence, as used herein, also includes machine learning, deep learning and the like.
25 Furthermore, as used herein the term images refers to pictures and video.

[0003] Referring to FIG. 1, a method of image capture and processing for use in artificial intelligence tasks, according to the convention art, is shown. Images are generally captured in a given format, such as frames of Red, Green and Blue (RGB) pixel data, referred to herein as the native format 110. The native format can comprise a relatively
30 large amount of data that consumes a relatively large amount of communication bandwidth and or a relatively large amount of memory space to transfer and or store the image data in

the native format. Therefore, the native format image data is typically converted to a transformed domain data format 120. For example, RGB pixel data can be converted into YCbCr discrete cosine transform (DCT) encoded components for transmission and storage. The transformed domain image data can then be transmitted and or stored 130 for subsequent processing by artificial intelligence tasks. Artificial intelligence tasks, however, typically process image data in a native format such as RGB pixel data. Therefore, the transformed domain data format is typically converted back to the native data format 140, before processing in artificial intelligence tasks 150.

[0004] Referring now to FIG. 2, an exemplary method of converting native format image data to transformed domain image data 120 according to the conventional art is shown. As illustrated, a frame of RGB image data 210 can include an array of red pixel values 212, an array of green pixel values 214 and an array of blue pixel values 216. Each value can be represented by eight bits. For example, frames of pixel data in the Red, Green and Blue (RGB) color space 210 can be converted 220 into frames of image data in the YCbCr luma-chroma color space 230. The frame of YCbCr image data 230 can include a luma component 232, a blue-difference chroma component 234 and a red-difference chroma component 236. The frames of YCbCr image data 230 can be domain transformed 240 utilizing a Discrete Cosine Transform (DCT) into components of discrete cosine transformed YCbCr image data 250. The discrete cosine transformed YCbCr image data 250 can include a discrete cosine transformed luma component 252, a discrete cosine transformed blue-difference chroma component 254, and a discrete cosine transformed red-difference chroma component 256. The above exemplary method of converting native format image data to transformed domain image data is merely for illustrative purposes. It is to be appreciated that there are other color spaces and or other transformed domains that can be applied.

[0005] Referring now to FIG. 3, an exemplary method of converting transformed domain image data to native format image data 140 according to the conventional art is shown. For example, discrete cosine transformed YCbCr image data 310 can be inverse domain transformed 320 into frames YCbCr image data 330. The frames YCbCr image data 330 can be converted 340 into frames of RGB image data 350.

[0006] Image capture in a native data format, conversion of the native formatted image data to a transformed domain format for transmission and storage, and then

conversion back to the native format for processing in artificial intelligence tasks can also consume large amount of communication and processing resources. The increased communication and processing load can result in increased processing latency and or increased power consumption. Therefore, there is a continuing need for improved image
5 capture and processing by Artificial Intelligence, machine learning, or deep learning tasks.

SUMMARY OF THE INVENTION

[0007] The present technology may best be understood by referring to the following description and accompanying drawings that are used to illustrate embodiments of the present technology directed toward techniques to dynamically gate encoded image
10 components utilized in Artificial Intelligence (AI) tasks.

[0008] In one embodiment, a method of computing an artificial intelligence task can include receiving components of transformed domain image data. The components of transformed domain image data can be processed in accordance with an artificial intelligence task based on one or more gating flags. In one implementation, the one or more
15 gating flags can be based on available computations resources. In another implementation, the one or more gating flags can be based on the importance of the components of transformed domain image data. In yet another implementation, the one or more gating flags can be based on the combination of the importance of the components of transformed domain image data and the computations resources available for processing the components
20 of transformed domain image data.

[0009] In another embodiment, one or more computing device executable instructions stored in one or computing device readable media (e.g., memory) that when executed by one or more compute unit (e.g., processors) can perform a method of computing an artificial intelligence task. The method can include receiving components of transformed domain
25 image data. The components of transformed domain image data can be processed based on one or more gating flags. The one or more gating flags can be based on one or more indications of an importance of the components of transformed domain image data, one or more indications of computations resources available, or similar factors, and combinations thereof.

30 [0010] This Summary is provided to introduce a selection of concepts in a simplified form that are further described below in the Detailed Description. This Summary is not

intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used to limit the scope of the claimed subject matter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] Embodiments of the present technology are illustrated by way of example and not by way of limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements and in which:

FIG. 1 shows a method of image capture and processing for use in artificial intelligence tasks, according to the convention art.

FIG. 2 shows an exemplary method of converting native format image data to transformed domain image data according to the conventional art.

FIG. 3 shows an exemplary method of converting transformed domain image data to native format image data according to the conventional art.

FIG. 4 shows a system for determining importance of encoded image components for artificial intelligence tasks, in accordance with aspects of the present technology.

FIG. 5 shows an exemplary processor core, in accordance with aspects of the present technology.

FIG. 6 shows a method of determining importance of encoded imaged components for artificial intelligence tasks, in accordance with aspects of the present technology.

FIG. 7 shows a system for determining importance of encoded image components for artificial intelligence tasks, in accordance with aspects of the present technology.

FIG. 8 illustrates exemplary components of the transformed domain image data, in accordance with aspects of the present technology.

FIG. 9 illustrates relative importance of exemplary components of the transformed domain image data, in accordance with aspects of the present technology.

FIG. 10 shows a method of computing an artificial intelligence task, in accordance with aspects of the present technology.

FIG. 11 shows a processor for computing the artificial intelligence task, in accordance with aspects of the present technology.

DETAILED DESCRIPTION OF THE INVENTION

[0012] Reference will now be made in detail to the embodiments of the present technology, examples of which are illustrated in the accompanying drawings. While the present technology will be described in conjunction with these embodiments, it will be understood that they are not intended to limit the invention to these embodiments. On the contrary, the invention is intended to cover alternatives, modifications and equivalents, which may be included within the scope of the invention as defined by the appended claims. Furthermore, in the following detailed description of the present technology, numerous specific details are set forth in order to provide a thorough understanding of the present technology. However, it is understood that the present technology may be practiced without these specific details. In other instances, well-known methods, procedures, components, and circuits have not been described in detail as not to unnecessarily obscure aspects of the present technology.

[0013] Some embodiments of the present technology which follow are presented in terms of routines, modules, logic blocks, and other symbolic representations of operations on data within one or more electronic devices. The descriptions and representations are the means used by those skilled in the art to most effectively convey the substance of their work to others skilled in the art. A routine, module, logic block and/or the like, is herein, and generally, conceived to be a self-consistent sequence of processes or instructions leading to a desired result. The processes are those including physical manipulations of physical quantities. Usually, though not necessarily, these physical manipulations take the form of electric or magnetic signals capable of being stored, transferred, compared and otherwise manipulated in an electronic device. For reasons of convenience, and with reference to common usage, these signals are referred to as data, bits, values, elements, symbols, characters, terms, numbers, strings, and/or the like with reference to embodiments of the present technology.

[0014] It should be borne in mind, however, that all of these terms are to be interpreted as referencing physical manipulations and quantities and are merely convenient labels and are to be interpreted further in view of terms commonly used in the art. Unless specifically stated otherwise as apparent from the following discussion, it is understood that through discussions of the present technology, discussions utilizing the terms such as "receiving," and/or the like, refer to the actions and processes of an electronic device such

as an electronic computing device that manipulates and transforms data. The data is represented as physical (e.g., electronic) quantities within the electronic device's logic circuits, registers, memories and/or the like, and is transformed into other data similarly represented as physical quantities within the electronic device.

5 [0015] In this application, the use of the disjunctive is intended to include the conjunctive. The use of definite or indefinite articles is not intended to indicate cardinality. In particular, a reference to "the" object or "a" object is intended to denote also one of a possible plurality of such objects. It is also to be understood that the phraseology and terminology used herein is for the purpose of description and should not be regarded as
10 limiting.

 [0016] Referring now to FIG 4, a system for determining importance of encoded image components for Artificial Intelligence (AI) tasks, in accordance with aspects of the present technology, is shown. The system can include one or more image capture units 402 or image storage units 404 and one or more processors 406 communicatively coupled by
15 one or more communication interfaces 408-414. The one or more communication interfaces 408-414 can include one or more wired network communication links, one or more wireless network communication links, one or more data buses, or the like, or combinations thereof.

 [0017] In one implementation, the one or more image capture units 402 can be a camera, video camera or the like for generating one or more frames of image data in a
20 native format. The one or more image storage units 404 can be a hard disk drive (HDD), solid state storage device (SSD), random access memory (RAM), flash memory, network attached storage (NAS), or the like, or combinations thereof, for storing the components of transformed domain image data. The one or more image capture units 402 and or one or more image storage units 404 can include encoder circuitry 416 to convert the image data in
25 the native format to components of transformed domain image data. Alternatively or in addition, the encoder circuitry 416 can be separate from the one or more capture units 402 and or the one or more image storage units 404. In one implementation, the encoder circuitry 416 can include a discrete cosine transform engine 418, a quantization engine 420 and an entropy coding engine 422. The discrete cosine transform engine 418, quantization
30 engine 420 and entropy coding engine 422 can be configured to convert native format image data to a transformed domain data format. A detailed understanding of the discrete cosine transform engine 418, the quantization engine 420 and the entropy coding engine

422 are not necessary for an understanding of aspects of the present technology and therefore will not be discussed further herein.

[0018] The one or more processors 406 can include one or more central processing units (CPUs), one or more cores of one or more central processing units (CPUs), one or more graphics processing units (GPUs), one or more cores of one or more graphics processing units (GPUs), one or more neural processing units (NPU), one or more cores of one or more neural processing units (NPU), one or more vector processors, one or more memory processing units, or the like, or combinations thereof. In one implementation, the one or more processors 406 can be one or more neural processing units. An neural processing unit can include one or more communication interfaces, such as peripheral component interface (PCIe4) 424 and inter-integrated circuit (I²C) interface 426, an on-chip circuit tester, such as a joint test action group (JTAG) engine, a direct memory access engine 428, a command processor (CP) 430, and one or more cores 432-438. The one or more cores 432-438 can be coupled in a single-direction ring bus configuration. The one or more cores 432-438 can execute one or more sets of computing device executable instructions to perform one or more functions, such as entropy decoding 440, component importance determining 442, artificial intelligence tasks 444 and or the like. One or more functions can be performed on an individual core 432-438, can be distributed across a plurality of cores 432-438, can be performed along with one or more other functions on one or more cores, and or the like.

[0019] In one implementation, the one or more processors 406 can be implemented in one or more computing devices 446. The one or more computing devices 446 can be, but are not limited to, cloud computing platforms, edge computing devices, servers, workstations, personal computers (PCs).

[0020] Referring now to FIG. 5, an exemplary processor core, in accordance with aspects of the present technology, is shown. In one implementation, the processing core 500 can be a neural processing unit core. The processing core 500 can include a tensor engine (TE) 510, a pooling engine (PE) 515, a memory copy engine (ME) 520, a sequencer (SEQ) 525, an instructions buffer (IB) 530, a local memory (LM) 535, and a constant buffer (CB) 540. The local memory 535 can be pre-installed with model weights and can store in-use activations on-the-fly. The constant buffer 540 can store constant for batch normalization, quantization and the like. The tensor engine 510 can be utilized to accelerate

fused convolution and or matrix multiplication. The pooling engine 515 can support pooling, interpolation, region-of-interest and the like operations. The memory copy engine 520 can be configured for inter- and or intra-core data copy, matrix transposition and the like. The tensor engine 510, pooling engine 515 and memory copy engine 520 can run in parallel. The sequencer 525 can orchestrate the operation of the tensor engine 510, the pooling engine 515, the memory copy engine 520, the local memory 535, and the constant buffer 540 according to instructions from the instruction buffer 530. The neural processing unit core 500 can provide convolution efficient computation under the control of operation fused coarse grained instructions for functions such as component importance determining 442, artificial intelligence tasks 444 and or the like. A detailed description of the exemplary neural processing unit core 500 is not necessary to an understanding of aspects of the present technology, and therefore will not be described further herein.

[0021] The system for determining the importance of encoded image components for artificial intelligence task will be further explained with reference to FIG. 6, along with FIG. 4 as described above. FIG. 6 shows a method of determining the importance of encoded imaged components. The method may be implemented as computing device-executable instructions (e.g., computer program) that are stored in computing device-readable media (e.g., computer memory) and executed by a computing device (e.g., processor). The method can include receiving components of transformed domain image data, at 610. The components of transformed domain image data can be received by the one or more processors 406 from the one or more image capture units 402 or one or more storage units 404 across the one or more communication interfaces 408-414. In one implementation, the components of transformed domain image data can be one or more frames of Discrete Cosine Transform (DCT) YCbCr image data components. In other implementations, the component of transformed domain image data can be components of frequency domain image data, components of Fourier Transform (FT) image data, components of Wavelet Transform image data, components of Discrete Wavelet Transform (DWT) image data, components of Hadamard transform image data, or components of Walsh transform image data, or the like. In one implementation, the received components of transformed domain image data can be generated from native format image data as described above with reference to FIG. 2.

[0022] At 620, the relative importance of the components of transformed domain image data can be determined by a component importance determining function 442 on the one or more processors 406 for a given artificial intelligence tasks. As previously stated, the use of the term Artificial Intelligence herein is intended to also includes machine learning, deep learning and the like. In one implementation, determining the relative importance of the components of the transformed domain image data can include gating the components of the transformed domain image data to turn on select ones of the components of the transformed domain image data for input to a Deep Neural Network (DNN) to determine the relative importance of the component of the transformed domain image data. In another implementation, determining the relative importance of the components of the transformed domain image data can include gating the components of the transformed domain image data based on a cost function to control select ones of the components of the transformed domain image data for input to a Deep Neural Network (DNN) to determine the relative importance of the component of the transformed domain image data. For example, the cost function can include a first term based on an error between a prediction and a target value, and a second term based on the number of active channels in accordance with Equation 1:

$$Cost = Loss(prediction, target) + \lambda \#Proportion(active - channels) \quad (1)$$

In yet another implementation, determining the relative importance of the components of the transformed domain image data can include gating the components of the transformed domain image data in which more important components is more likely to be turned on than less important components for input to a Deep Neural Network (DNN) to determine the relative importance of the component of the transformed domain image data.

[0023] At 630, an indication of the relative importance of the components of the transformed domain image data can be output. In one implementation, the indication of the relative importance of the components of the transformed domain image data can be provided to the given artificial intelligence task for use in performing the artificial intelligence task on the components of the transformed domain image data.

[0024] Referring now to FIG. 7, a system for determining importance of encoded image components for artificial intelligence tasks, in accordance with aspects of the present technology, is shown. The system comprises a plurality of neural networks 702-710 configured to receive a corresponding ones of a plurality of transformed domain image data components 712-720. In one implementation, the transformed domain image data components can include 64 discrete cosine transformed luma Y components, 64 discrete cosine transformed blue-difference chroma Cb components, and 64 discrete cosine transformed blue-difference chroma Cr components.

[0025] The plurality of neural networks 702-710 can be configured to predict if the corresponding one of the plurality of transformed domain image data components 712-720 comprises an important component for an artificial intelligence task. If the transformed domain image data component is an important component, the corresponding neural network can generate a first indication. If the transformed domain data component is not an important component, the corresponding neural network can generate a second indication. In one implementation, the respective neural networks generate an indication of '1' when the corresponding transformed domain data component is important, and an indication of '0' when the corresponding transformed domain data component is not important for a given artificial intelligence task.

[0026] The plurality of transformed domain image data components 712-720 can be gated 722-730 in accordance with the corresponding indication generated by the corresponding neural network 702-710. In one implementation, the plurality of transformed domain image data components 712-720 can be gated 722-730 by multiplying the corresponding indication from the corresponding neural networks 702-710. If the indication is a '1,' the received transformed domain image data component 712, 718, 720 can be provided as in important transformed domain image data component 732, 738, 740. If the indication is a '0,' the received transformed domain image data component 714, 716 can be blocked 734, 736. The indication of the importance of the transformed domain image data components is not limited to values of '1' and '0'. In other instances, the indication can be a value between 0 and 1, wherein the higher the value the greater the importance of the transformed domain image data component.

[0027] Referring now to FIG. 8, exemplary components of the transformed domain image data, in accordance with aspects of the present technology, are illustrated. The

exemplary components of transformed domain image data may be based on the ImageNet dataset. ImageNet includes more than 14 million images in more than 20,000 categories that have been annotated to indicate what objects are pictured. The neural network used to determine the importance of the transformed domain image data components can be the MobileNet_v2, which is a convolution neural network that has been trained on more than a million images from the ImageNet dataset. The luma Y components of the discrete cosine transformed image data can include an 8x8 matrix of 64 channels. Similarly, the blue-difference chroma Cb components of discrete cosine transformed image data can include an 8x8 matrix of 64 channels, and the red-difference chroma Cr components of discrete cosine transformed image data can include an 8x8 matrix of 64 channels. A value of the relative importance of the transformed domain image data components can be assigned to each channel value. In the illustrated example, channels 0-3, 8-11, 16-19 and 24-26 in the discrete cosine transformed luma Y components can have importance values greater than '0.' Channels 0, 1 and 8 in the discrete cosine transformed blue-difference chroma, and channels 0, 1, 8 and 9 in the discrete cosine transformed red-difference chroma components can have importance values greater than '0.'

[0028] Referring now to FIG. 9, the relative importance of exemplary components of the transformed domain image data, in accordance with aspects of the present technology, are illustrated. The active channels, in the illustrated exemplary components of the discrete cosine transformed image data are mostly on luma Y channels 910, and partly on blue-difference chroma Cb channels 920, and red-difference chroma Cr channels 930. For the exemplary components of transformed domain image data based on the ImageNet dataset processed by the MobileNet_v2, using all 192 channels has been found to results in an accuracy of approximately 70%. In comparison, using 22 channels that are determined to be important can achieve an accuracy of approximately 69.3%. Accordingly, substantially similar accuracy can be achieved, while reducing the number of processed channels by 170.

[0029] Referring now to FIG. 10, method of computing an artificial intelligence task, in accordance with aspects of the present technology, is shown. The method of computing an artificial intelligence task will be further explained with reference to FIG. 11, which shows a processor for computing the artificial intelligence task. In one implementation, the processor 1100 can be one or more central processing units (CPUs), one or more cores of one or more central processing units (CPUs), one or more graphics processing units (GPUs),

one or more cores of one or more graphics processing units (GPUs), one or more artificial intelligence accelerators, one or more vector processors, one or more memory processing units, or the like, or combinations thereof. In one implementation, the processor 1100 can include one or more compute units 1110, one or more controllers 1115, and integral
5 memory 1120. In other implementations, the memory 1120 can be external to the processor 1100.

[0030] The method can include receiving components of transformed domain image data, at 1010. The components of transformed domain image data can be received by the memory 1120 of the processor 1100. In one implementation, the components of
10 transformed domain image data can be received from one or more image capture or storage units. In one implementation, the received components of transformed domain image data can be generated from native format image data as described above with reference to FIG. 2. In one implementation, the components of transformed domain image data can be one or more frames of Discrete Cosine Transform (DCT) YCbCr image data components. In other
15 implementations, the component of transformed domain image data can be components of frequency domain image data, components of Fourier Transform (FT) image data, components of Wavelet Transform image data, components of Discrete Wavelet Transform (DWT) image data, components of Hadamard transform image data, or components of Walsh transform image data, or the like.

20 [0031] At 1020, an artificial intelligence task can be configured to process the components of the transformed domain image data based on one or more gating flags. In one implementation, the one or more gating flags can be arranged in a vector of bits indicating corresponding flag states or values. For example, the vector of bits can indicate one or more on components of the transformed domain image data and one or more off
25 components of the transformed domain image data. The vector bits can be read to accelerate hardware by reading and or processing turned on components of the transformed domain image data and skipping turned off components. In one implementation, the one or more gating flags can be based on one or more indications of available computational resources 1030. For example, when the current utilization of the processor 1100 is
30 relatively low (e.g., below a threshold) the one or more gating flags can be configured for processing a larger subset of the components of transformed domain image data by the processor 1100. When the current utilization of the processor 1100 is relatively high (e.g.,

above a threshold) the one or more gating flags can be configured to for processing a smaller subset of the components of transformed domain image data. In another implementation, the one or more gating flags can be based on one or more indications of the relative importance of the components of transformed domain image data 1040. For example, the gating flags can be configured for processing a subset of the transformed domain image data components determined to be important by the processor 1100 for the given artificial intelligence task. In yet another implementation, the one or more gating flags can be based on the one or more indications of the relative importance of the components of transformed domain image data 940 and the available computational resources 1030. For example, when the current utilization of the processor 1100 is relatively low (e.g., below a threshold), the one or more gating flags can be configured for processing a larger subset of the important transformed domain image data components. When the current utilization of the processor 1100 is relatively high (e.g., above a threshold), the one or more gating flags can be configured for processing a smaller subset of important transformed domain image data components. The one or more gating flags can be dynamically adjusted based on changes in the available computational resources and or the importance of components of the transformed image data. Furthermore, the gating flags are not limited to representing available computational resource and or the importance of the transformed image data components. The gating flags can also represent any number of other factors related to the processing of the components of transformed domain image data for an artificial intelligence task.

[0032] In one implementation, the controller 1115 can include one or more channel switches 1125 and one or more micro-operation scheduler 1130. The one or more channel switches 1125 can be configured to receive the one or more gating flags computed by one or more compute units 1110. The one or more micro-operation controllers 1130 can be configured to receive instructions 1140, such as convolution commands, and generate corresponding micro-operations 1145 for execution by the one or more compute units 1110 to perform functions, such as convolution operations, for performing one or more artificial intelligence tasks. The one or more micro-operations schedulers 1030 can also be configured to generate addresses 1150 for transferring data, such as components of transformed domain image data 1155, into the memory 1120. The one or more micro-operations scheduler 1130 can also be configured to generate addresses 1150 for

transferring data, such as components of transformed domain image data 1155, for corresponding micro-operations to the one or more compute units 1110. The one or more channel switches 1125 can be configured to selectively control the generation of micro-operations 1145 and addresses 1150 by the one or more micro-op schedulers 1130 based on the one or more gating flags 1135 from the one or more compute units 1110. In one implementation, the one or more channel switches 1125 can cause the one or more micro-op scheduler 1130 to generate micro-operations 1145 to multiply the state of a corresponding gating flag with a corresponding transformed domain image data component. For example, if the corresponding gating flag is a '0' indicating that the component is not an important component and or there is not enough processing resources to utilize the component, the corresponding micro-operation 1145 can multiply the corresponding gating flag with the corresponding transformed domain image data component resulting in a corresponding transformed domain image data component of all zeros, making subsequent computations easier to perform. If the corresponding gating flag is a '1' indicating that the component is an important component and or there are enough processing resources, the corresponding transformed domain image data component is fed to subsequent computations unchanged. Alternatively, the one or more channel switches 1125 can cause the one or more micro-op schedulers 1130 to skip the generation of corresponding micro-operations 1145 and addresses 1150 when the corresponding gating flag is a "0" such that the one or more compute units 1110 do not have to perform operations with respect to the corresponding transformed domain image data component. After performing the received instructions 1140, such as the convolution commands of a given artificial intelligence task, the processor 1100 can output results of the artificial intelligence task, at 1050.

[0033] Accordingly, performing artificial intelligence tasks can be performed on a subset of transformed domain image data components based on one or more factors such as available computational resources, the importance of the components of the transformed domain image data, and the like. Artificial intelligence tasks can also be performed directly on the components of transformed domain image data, which can also advantageously reduce computational workload, data transmission bandwidth, processing latency, power consumption and or the like because the components of transformed domain image data does not need to be converted back to the native image data format before performing the artificial intelligence tasks. Performing artificial intelligence tasks only on a subset of the

components of transformed domain image data found to be more important than other components of the transformed domain image data can also further reduce computational workload, data transmission bandwidth, processing latency, power consumption and or the like. Performing artificial intelligence tasks only on the subset of the components of
5 transformed domain image data found to be more important than other components of the transformed domain image data has been found to achieve almost equal accuracy as compared to performing the artificial intelligence task on the native format image data.

[0034] The foregoing descriptions of specific embodiments of the present technology have been presented for purposes of illustration and description. They are not intended to
10 be exhaustive or to limit the invention to the precise forms disclosed, and obviously many modifications and variations are possible in light of the above teaching. The embodiments were chosen and described in order to best explain the principles of the present technology and its practical application, to thereby enable others skilled in the art to best utilize the present technology and various embodiments with various modifications as are suited to the
15 particular use contemplated. It is intended that the scope of the invention be defined by the claims appended hereto and their equivalents.

CLAIMS

What is claimed is:

1. A method comprising:
receiving components of transformed domain image data;
5 processing the components of transformed domain image data for an artificial intelligence task based on one or more gating flags; and
outputting results of the artificial intelligence task.
2. The method according to Claim 1, wherein the one or more gating flags are based on one or more factors including availability of one or more computational resources for
10 processing the components of the transformed domain image data for the artificial intelligence task.
3. The method according to Claim 1, wherein the important components of the transformed domain image data are more likely to be used than less important components of the transformed domain image data based on the one or more gating flags.
- 15 4. The method according to Claim 1, further comprising:
wherein the one or more gating flags comprise a vector of bits indicating one or more on components of the transformed domain image data and one or more off components of the transformed domain image data; and
accelerating hardware using the vector of bits to determine the components of the
20 transformed domain image data for processing by the artificial intelligence task on the hardware
5. The method according to Claim 1, wherein the one or more gating flags are based on the combination of one or more indications of an importance of the components of transformed domain image data and one or more indications of computations resources
25 available for processing the components of transformed domain image data for the artificial intelligence task.
6. The method according to Claim 1, wherein the components of transformed domain image data includes components of frequency domain image data, components of Fourier

Transform (FT) image data, components of Discrete Cosine Transform (DCT) image data, components of Wavelet Transform image data, components of Discrete Wavelet Transform (DWT) image data, components of Hadamard transform image data, or components of Walsh transform image data.

5 7. The method according to Claim 6, wherein the components of transformed domain image data comprises components of encoded YCbCr color space image data.

8. The method according to Claim 6, wherein receiving components of transformed domain image data includes;

receiving RGB color space image data;
10 converting the RGB color space image data to YCbCr color space image data; and
encoding the YCbCr color space image data into components of Discrete Cosine Transform (DCT) YCbCr image data.

9. One or more computing device executable instructions stored one or computing device readable media that when executed by one or more processors performs a method of:

15 receiving components of transformed domain image data;
gating the components of transformed domain image data for processing by an artificial intelligence task based on one or more gating flags; and
outputting results of the artificial intelligence task.

10 10. The one or more computing device executable instructions stored one or computing device readable media that when executed by one or more compute units perform the method according to Claim 9, further comprising determining the one or more gating flags based on one or more factors include a computational resource for processing by the artificial intelligence task.

25 11. The one or more computing device executable instructions stored one or computing device readable media that when executed by one or more compute units perform the method according to Claim 11, further comprising determining the one or more gating flags wherein important components of the transformed domain image data are more likely to be used than less important components of the transformed domain image data.

12. The one or more computing device executable instructions stored one or computing device readable media that when executed by one or more compute units perform the method according to Claim 9, wherein:

the one or more gating flags comprise a vector of bits indicating one or more on
5 components of the transformed domain image data and one or more off components of the transformed domain image data; and

accelerating hardware using the vector of bits to determine the components of the transformed domain image data for processing by the artificial intelligence task on the hardware.

10 13. The one or more computing device executable instructions stored one or computing device readable media that when executed by one or more compute units perform the method according to Claim 9, wherein the one or more processors comprise one or more central processing units (CPUs), one or more cores of one or more central processing units (CPUs), one or more graphics processing units (GPUs), one or more cores
15 of one or more graphics processing units (GPUs), one or more vector processors, or one or more memory processing units.

14. The one or more computing device executable instructions stored one or computing device readable media that when executed by one or more compute units perform the method according to Claim 9, wherein the one or more processors comprise:

20 one or more compute units configured to process components of transformed domain image data according to one or more micro-operations for an artificial intelligence task; and

one or more controllers including:

one or more micro-op schedulers; and

25 one or more channel switches configured to selectively control a transfer of the components of transformed domain image data to the one or more compute units based on one or more gating flags, and to selectively control generation of the one or more micro-operations by the one or more micro-op schedulers based on the one or more gating flags.

15. The one or more computing device executable instructions stored one or
30 computing device readable media that when executed by one or more compute unit performs the method according to Claim 9, wherein the one or more gating flags are based

on the combination of one or more indications of an importance of the components of transformed domain image data and available computations resources for processing the components of transformed domain image data for an artificial intelligence task.

16. A device comprising:

5 one or more compute units configured to process components of transformed domain image data according to one or more micro-operations for an artificial intelligence task; and

one or more controllers including;

one or more micro-op schedulers; and

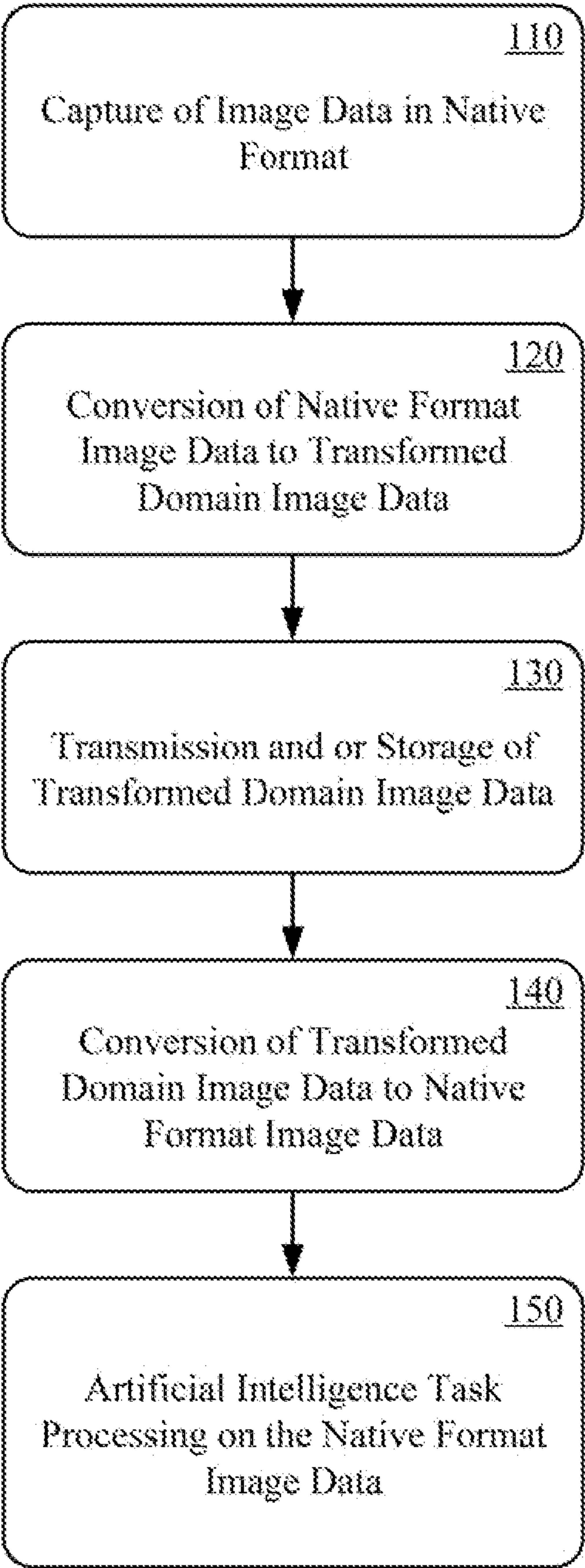
10 one or more channel switches configured to selectively control a transfer of the components of transformed domain image data to the one or more compute units based on one or more gating flags, and to selectively control generation of the one or more micro-operations by the one or more micro-op schedulers based on the one or more gating flags.

17. The device of Claim 16, wherein the one or more gating flags are based on one or
15 more indications of computations resources available for processing the components of transformed domain image data for the artificial intelligence task.

18. The device of Claim 16, wherein the one or more gating flags are based on one or more indications of an importance of the components of transformed domain image data.

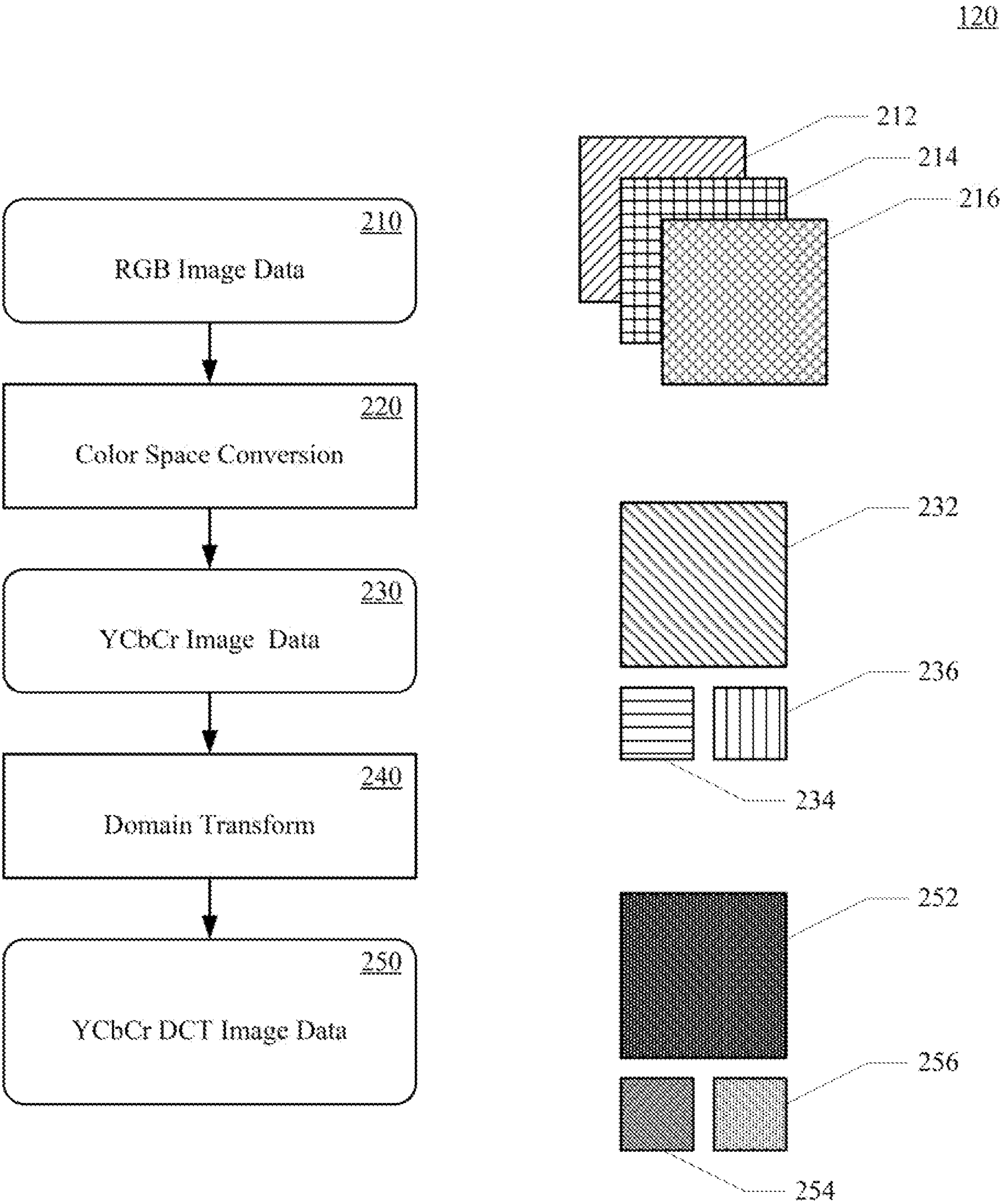
19. The device of Claim 16, wherein the one or more gating flags are based on the
20 combination of one or more indications of an importance of the components of transformed domain image data and one or more indications of computations resources available for processing the components of transformed domain image data for the artificial intelligence task.

20. The device of Claim 16, wherein the components of transformed domain image
25 data received by the one or more processor comprise one or more components of Discrete Cosine Transform (DCT) YCbCr image data.



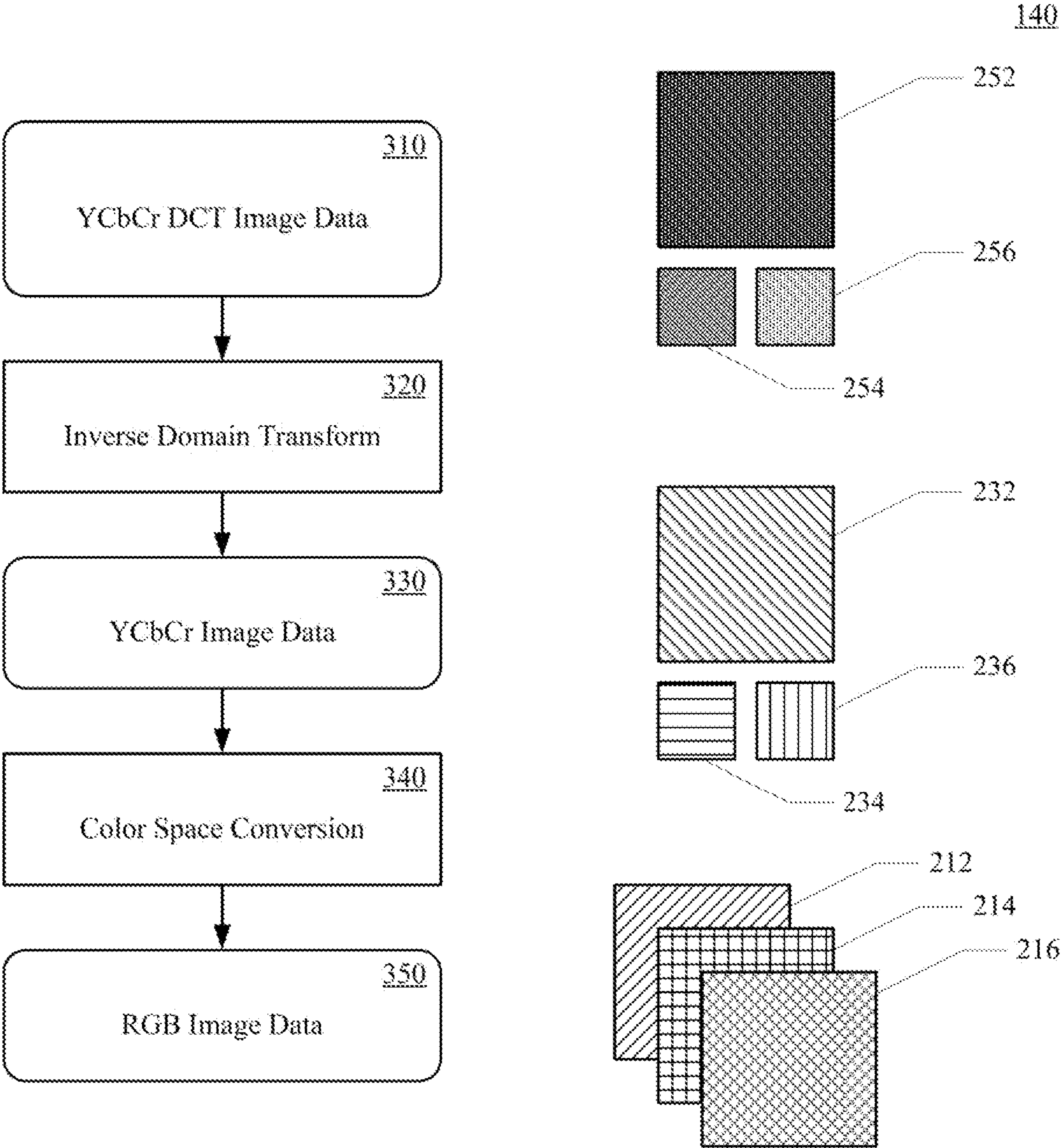
(Conventional Art)

FIG. 1



(Conventional Art)

FIG. 2



(Conventional Art)

FIG. 3

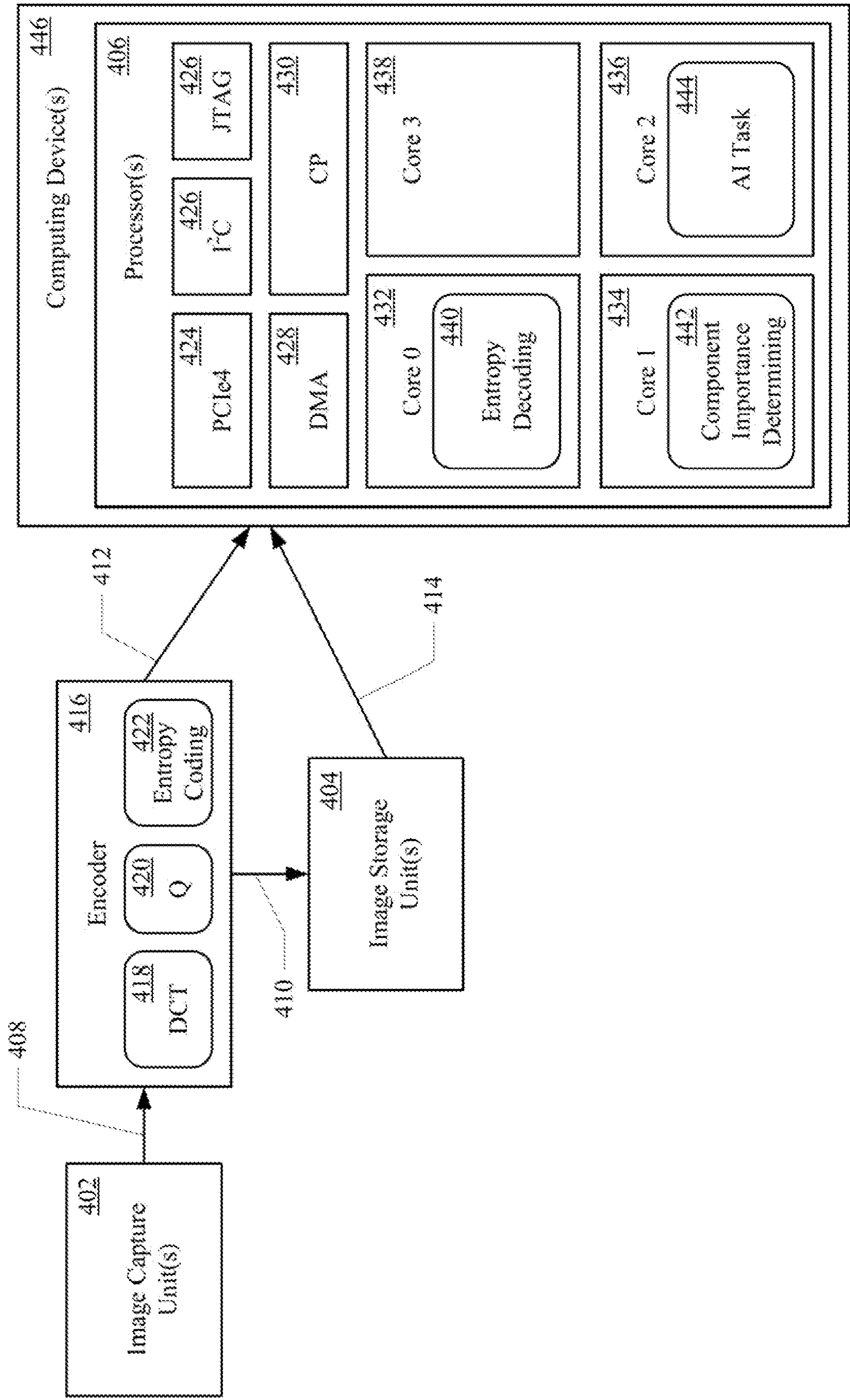


FIG. 4

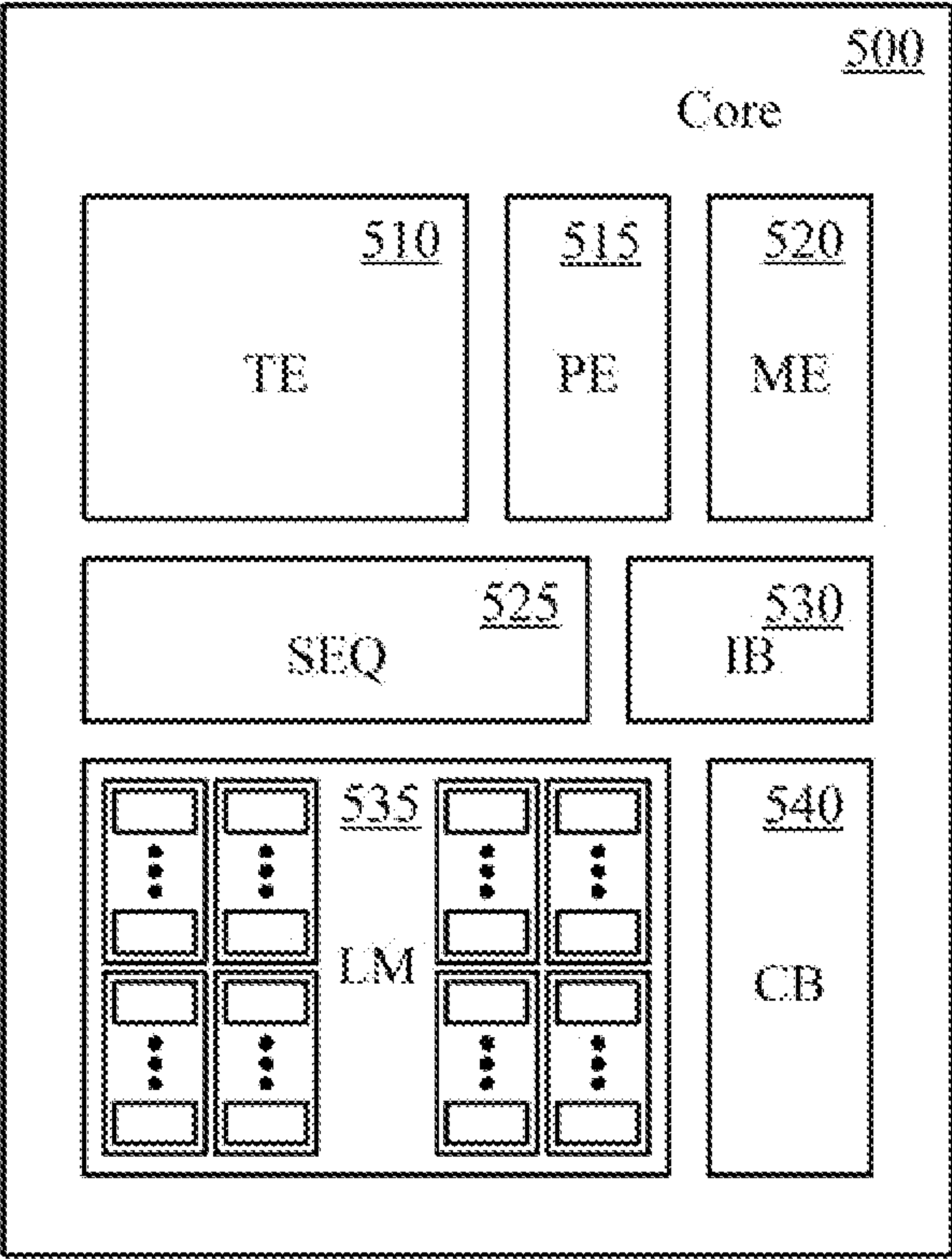


FIG. 5

6/11

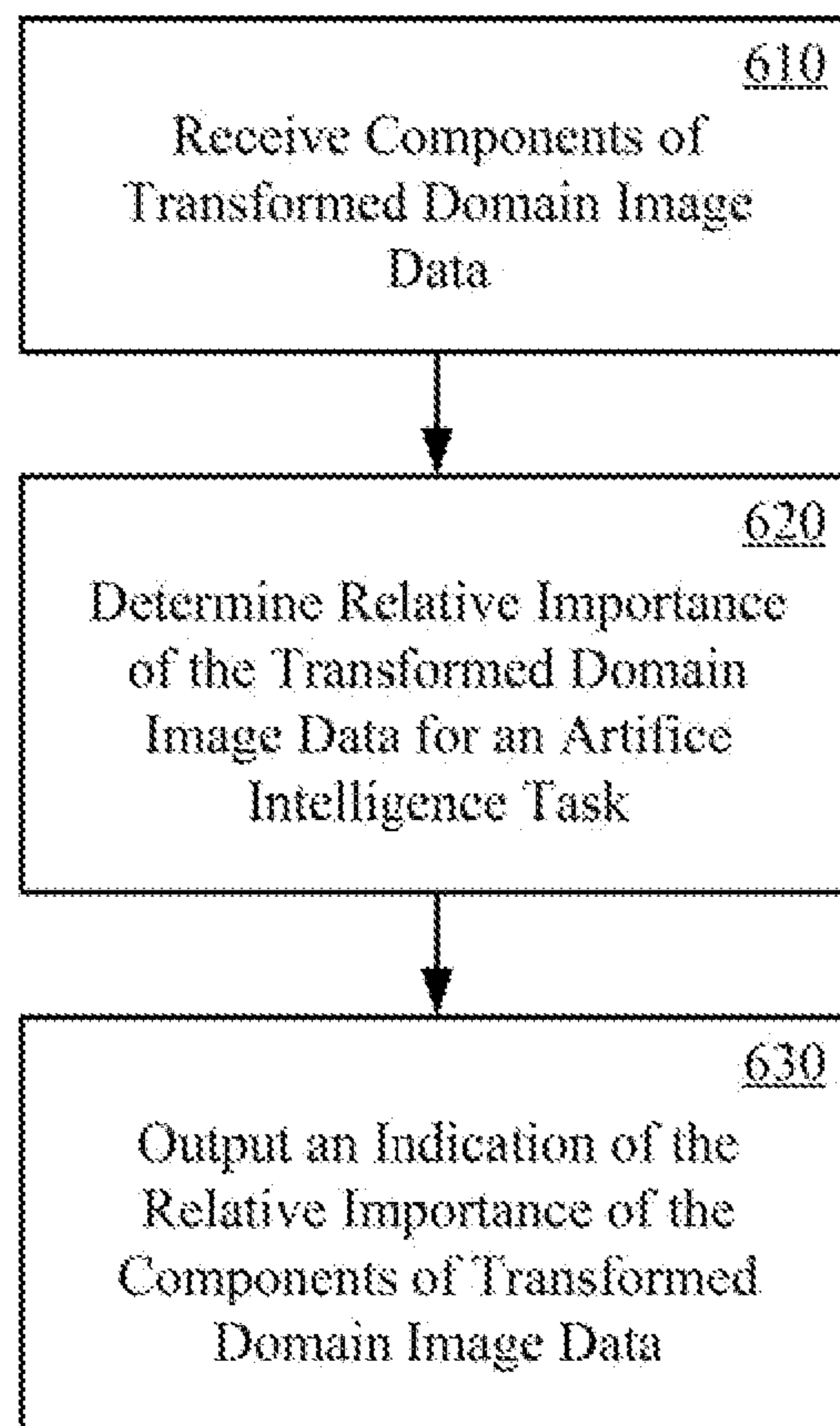


FIG. 6

7/11

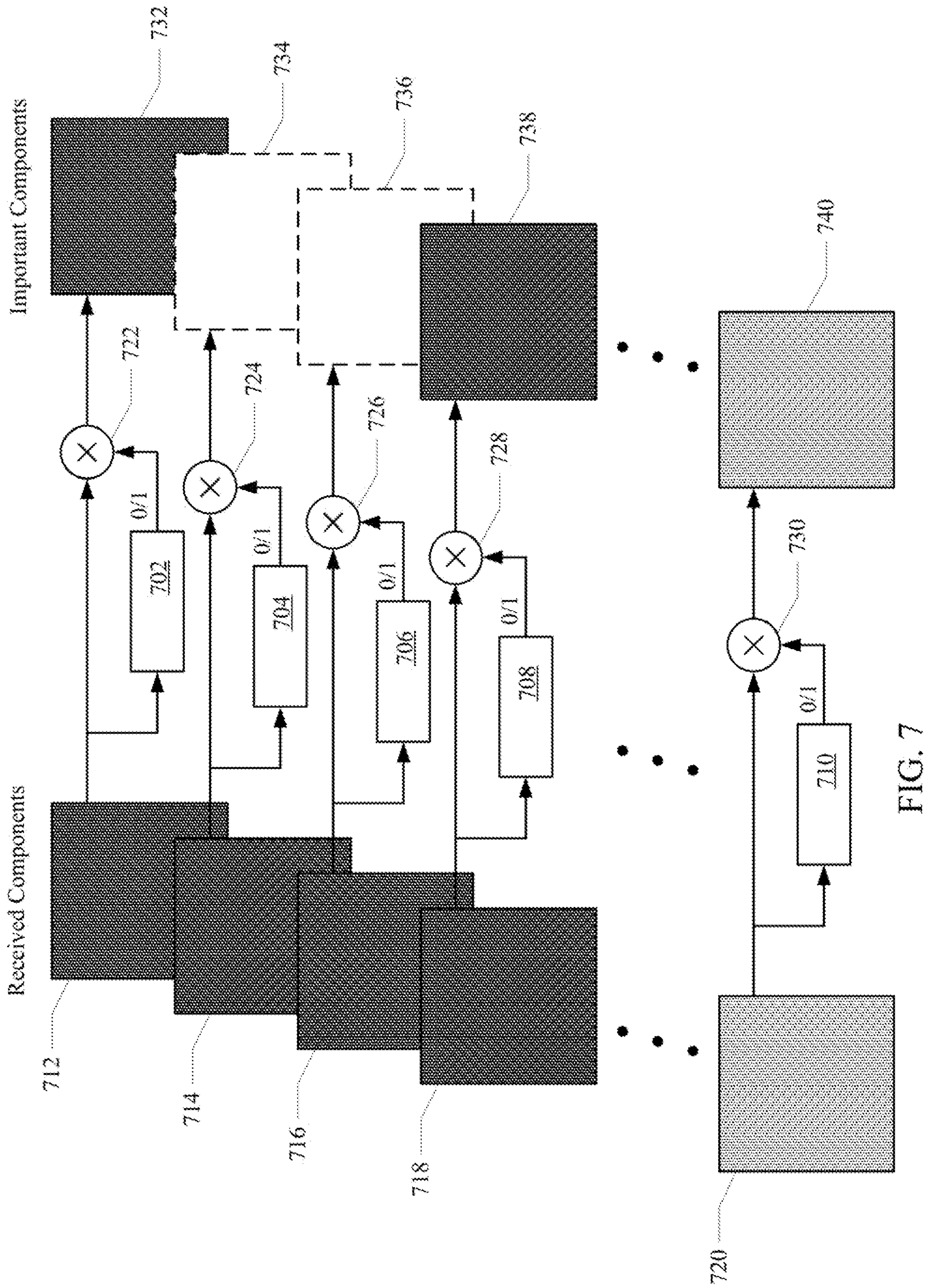


FIG. 7

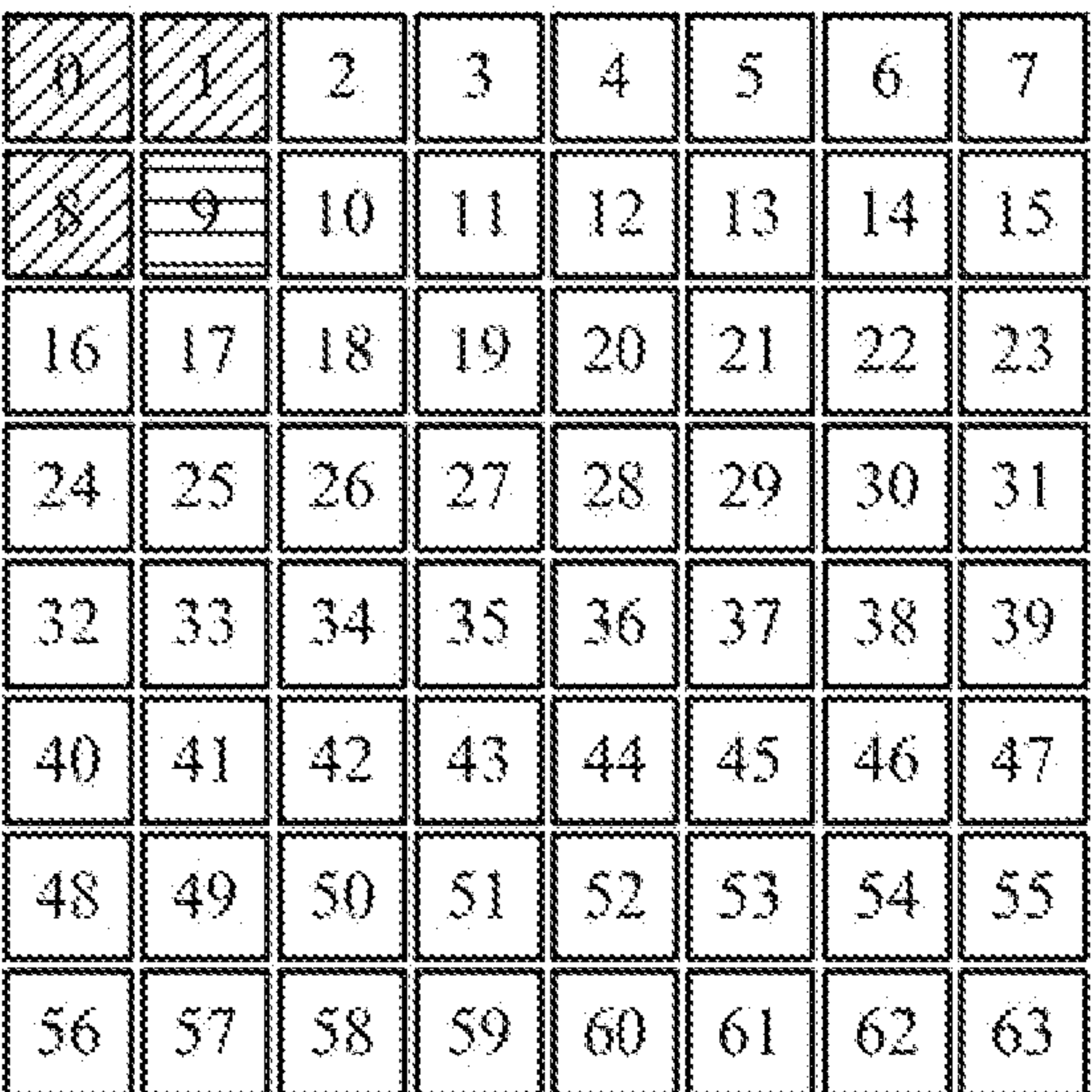
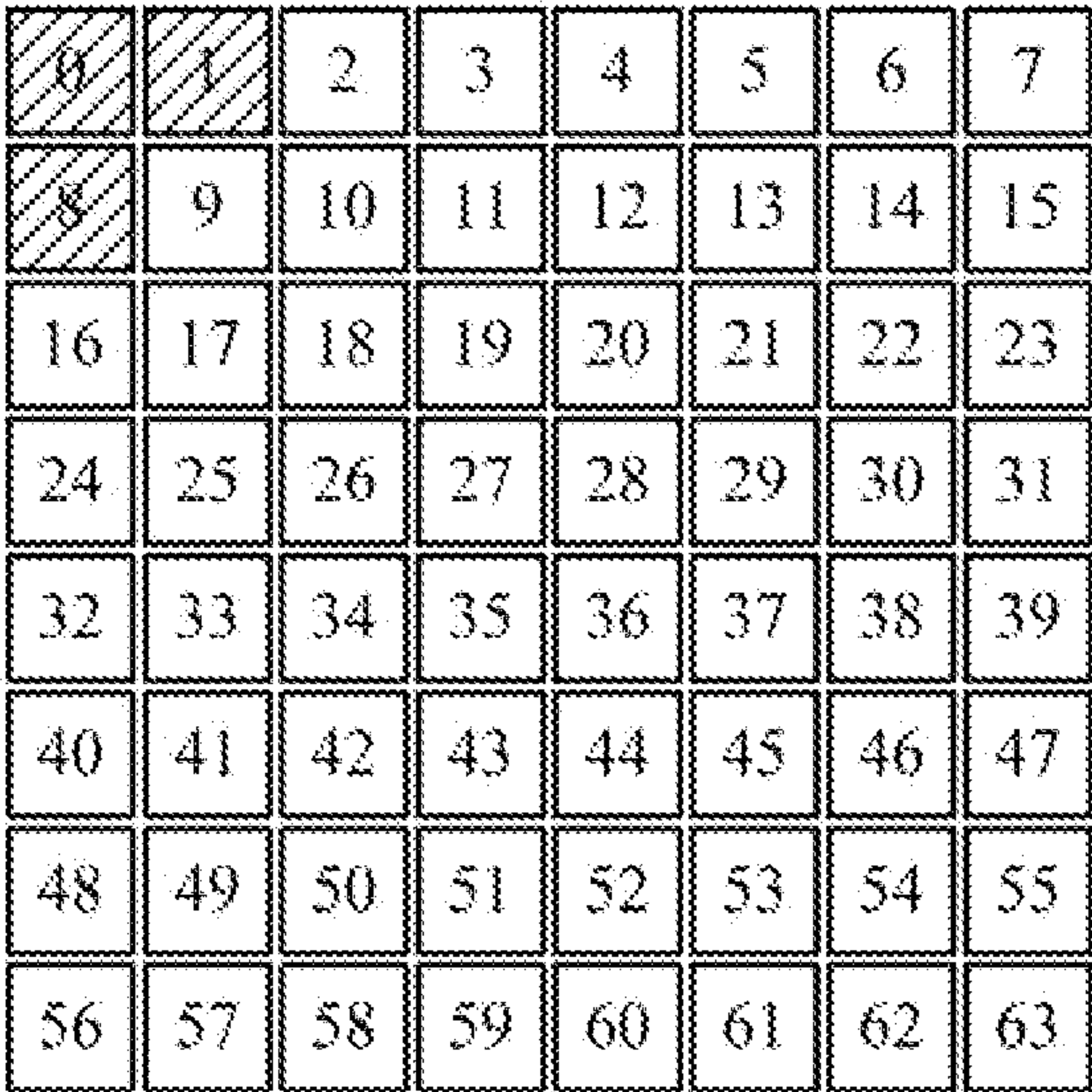
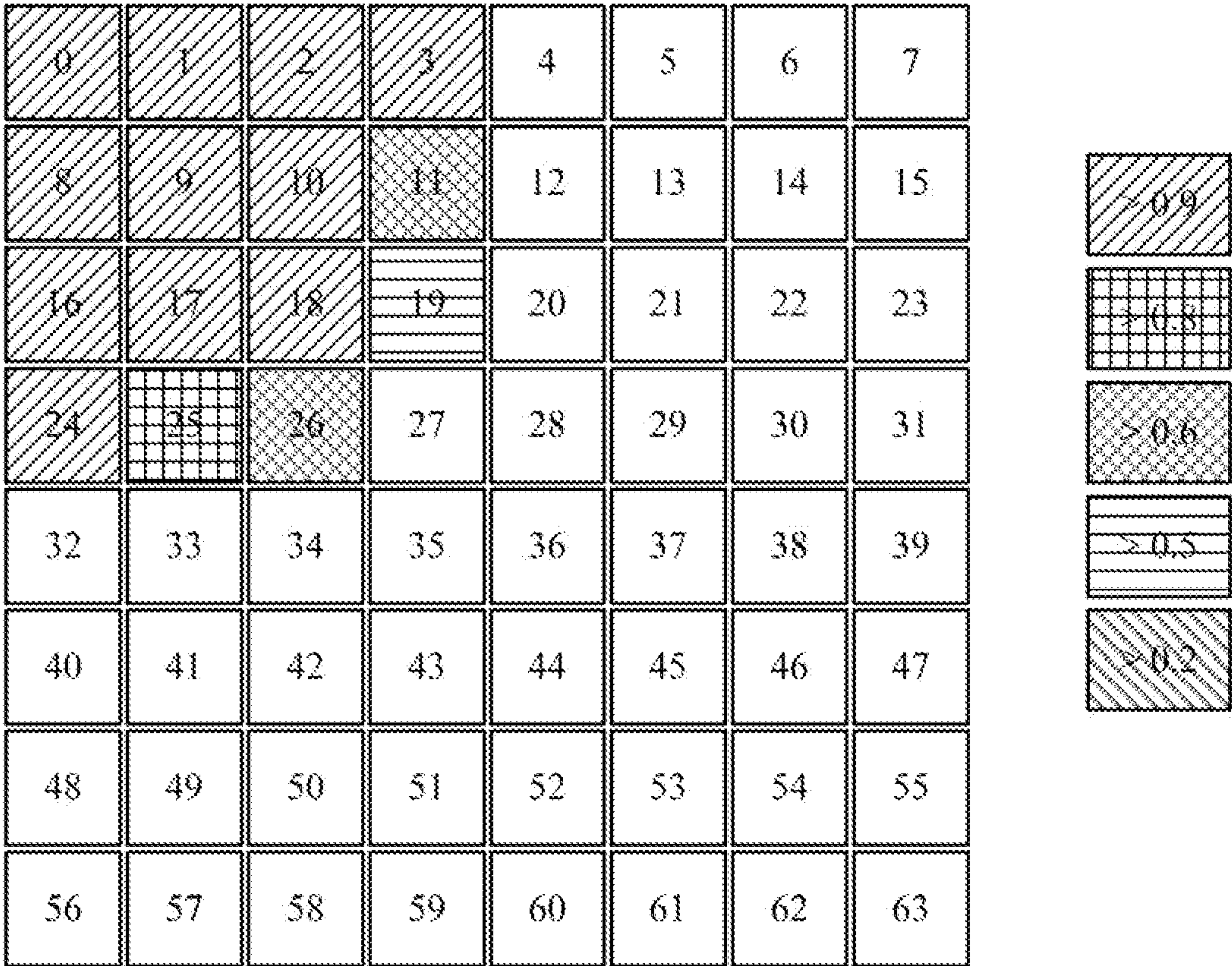


FIG. 8

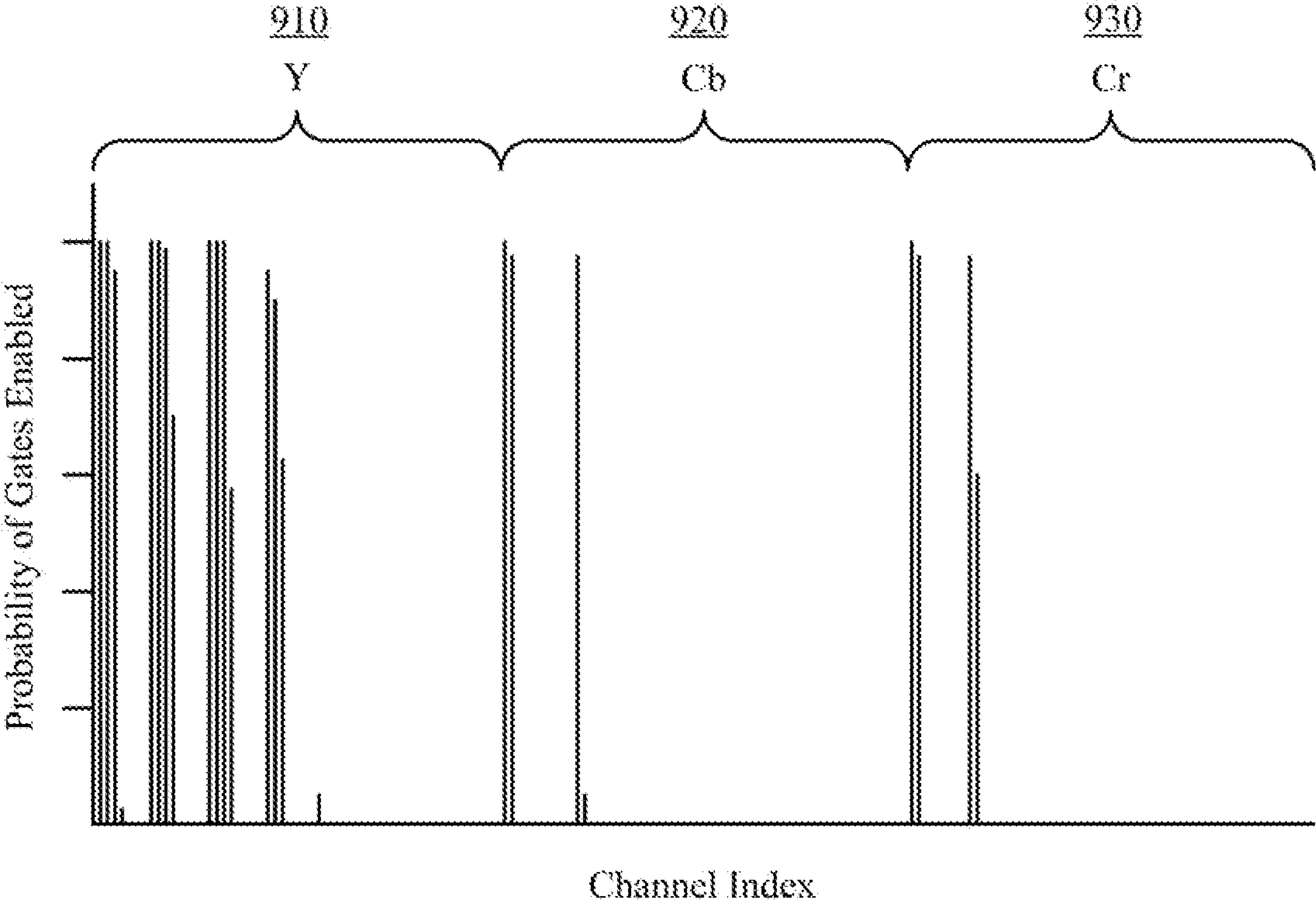


FIG. 9

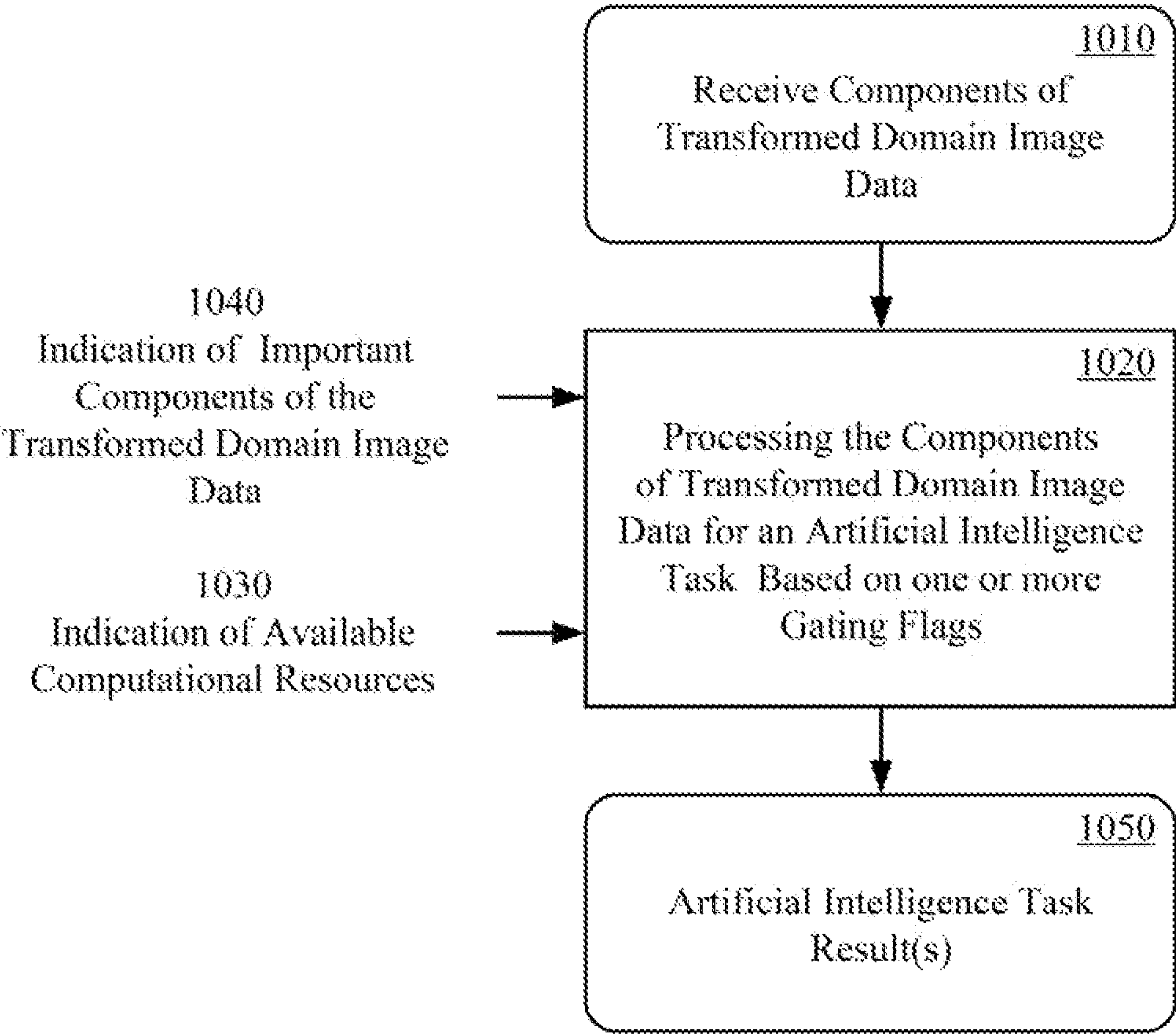


FIG. 10

1100

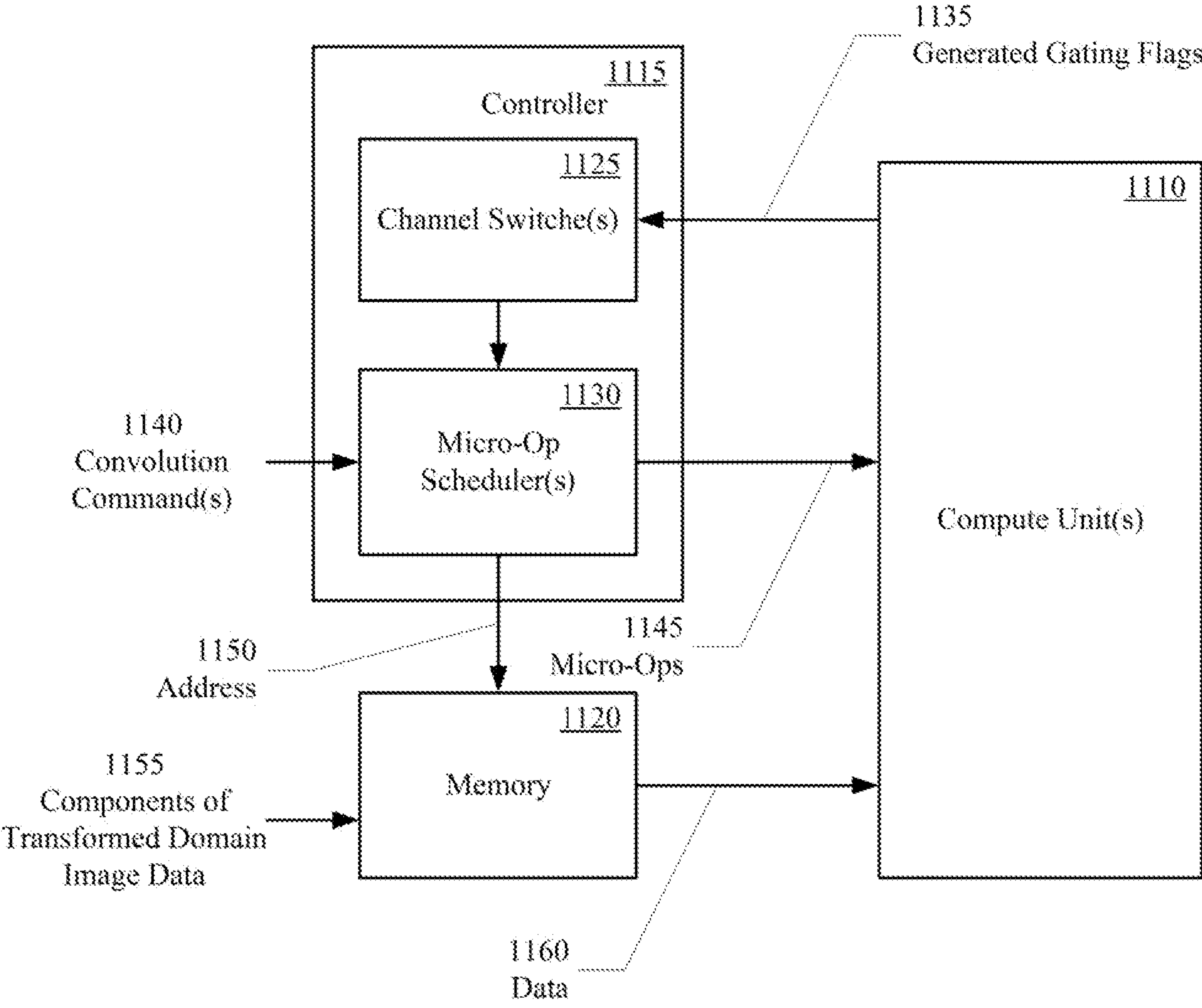


FIG. 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 20/60675

A. CLASSIFICATION OF SUBJECT MATTER

IPC - H04N 19/62 (2021.01)

CPC - H04N 19/62, H04N 19/18, H04N 19/46, H04N 19/124

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2016/0269733 A1 (Apple Inc.) 15 September 2016 (15.09.2016) entire document (especially para [0026], [0030], [0046], [0076]-[0081], [0091]).	1-20
Y	US 2019/0007685 A1 (Alcatel-Lucent) 03 January 2019 (03.01.2019) entire document (especially para [0042], [0060]-[0065], [0071]-[0077], [0086]-[0088], [0092], [0107]).	1-20

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

18 January 2021 (18.01.2021)

Date of mailing of the international search report

19 FEB 2021

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Lee Young

Telephone No. PCT Helpdesk: 571-272-4300