

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2025年2月27日(27.02.2025)



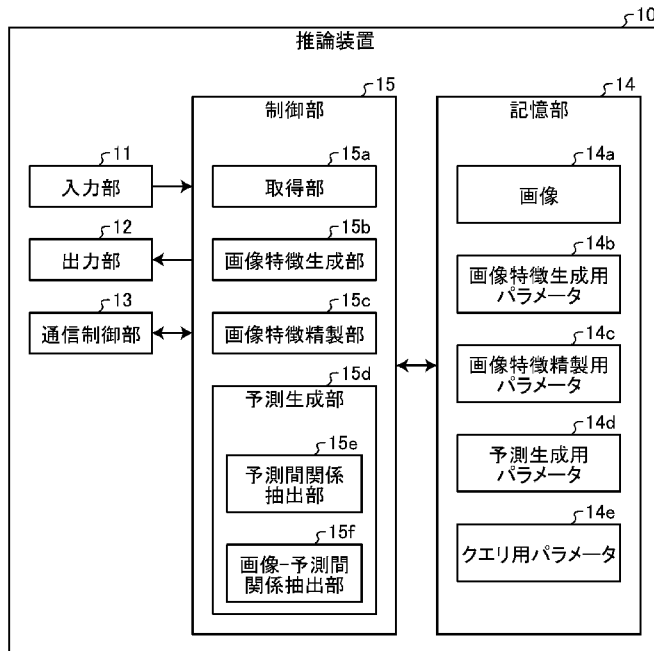
(10) 国際公開番号

WO 2025/041210 A1

- (51) 国際特許分類:
G06N 3/04 (2023.01) G06N 3/08 (2023.01)
- (21) 国際出願番号: PCT/JP2023/029898
- (22) 国際出願日: 2023年8月18日(18.08.2023)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (71) 出願人: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).
- (72) 発明者: 山根 大河(YAMANE, Taiga); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT 知的財産センタ内 Tokyo (JP). 鈴木 聡志(SUZUKI, Satoshi); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT 知的財産センタ内 Tokyo (JP).
- (74) 代理人: 弁理士法人酒井国際特許事務所 (SAKAI INTERNATIONAL PATENT OFFICE); 〒1000013 東京都千代田区霞が関3丁目8番1号 虎ノ門ダイビルイースト Tokyo (JP).
- (81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR,

(54) Title: INFERENCE DEVICE, INFERENCE METHOD, AND INFERENCE PROGRAM

(54) 発明の名称: 推論装置、推論方法、および推論プログラム



- 10 Inference device
11 Input unit
12 Output unit
13 Communication control unit
14 Storage unit
14a Image
14b Image feature generation parameters
14c Image feature refinement parameters
14d Prediction generation parameters
14e Query parameters
15 Control unit
15a Acquisition unit
15b Image feature generation unit
15c Image feature refinement unit
15d Prediction generation unit
15e Inter-prediction relationship extraction unit
15f Image-prediction relationship extraction unit

(57) Abstract: In an inference device (10), an acquisition unit (15a) acquires, for a plurality of tasks for a prescribed medium, a feature quantity of the medium and a query parameter representing a feature quantity of each task. A prediction generation unit (15d) predicts a task feature quantity of each task for the medium from the query parameters acquired by the acquisition unit (15a), using a first relationship between the query parameters representing the feature quantities of tasks that cross over different tasks and a second relationship between the feature quantity of the medium and the query parameter representing the feature quantity of each task.

HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, 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).

添付公開書類:

— 国際調査報告 (条約第21条(3))

(57) 要約: 推論装置 (10) において、取得部 (15a) が、所定のメディアに対する複数のタスクについて、メディアの特徴量と、各タスクの特徴量を表すクエリ用パラメータとを取得する。予測生成部 (15d) が、異なるタスクを横断した各タスクの特徴量を表すクエリ用パラメータの間の第1関係、およびメディアの特徴量と各タスクの特徴量を表すクエリ用パラメータとの第2関係を用いて、取得部 (15a) で取得されたクエリ用パラメータからメディアにおける各タスクのタスク特徴量を予測する。

明 細 書

発明の名称：推論装置、推論方法、および推論プログラム

技術分野

[0001] 本発明は、推論装置、推論方法、および推論プログラムに関する。

背景技術

[0002] 従来、画像、音声、自然言語等のメディアに対して、例えば画像に対する物体検出、領域分割、姿勢推定等の複数のタスクの推論結果が必要となる場合がある。複数のタスクに対応可能な推論手法として、例えば、Transformerと呼ばれる構造とObject Queryと呼ばれる予測の元となるパラメータを用いて物体検出を解くDETR (Detection Transformer) と呼ばれる手法が知られている（非特許文献1参照）。TransformerとObject Queryとを用いる手法は、物体検出の他にも姿勢推定や領域分割等の様々なタスクに有効であることが知られている。

先行技術文献

非特許文献

[0003] 非特許文献1：X. Zhu, “Deformable DETR: Deformable Transformers for End-to-End Object Detection”, 2021年

発明の概要

発明が解決しようとする課題

[0004] しかしながら、従来技術によれば、複数のタスクを同時に解くことは困難である。例えば、TransformerとObject Queryとを用いる手法は、様々なタスクに有効であるにも関わらず、各タスクに特化してタスク特化型モデルを作成して推論する必要がある。すなわち、タスクの数だけモデルを作成して推論を行うため、タスクの数に応じたコストがかかるという課題がある。

[0005] 本発明は、上記に鑑みてなされたものであって、複数のタスクの推論を効率よく実行可能とすることを目的とする。

課題を解決するための手段

[0006] 上述した課題を解決し、目的を達成するために、本発明に係る推論装置は、所定のメディアに対する複数のタスクについて、前記メディアの特徴量と、各タスクの特徴量を表すパラメータとを取得する取得部と、異なるタスクを横断した各タスクの特徴量を表すパラメータの間の第1関係、および前記メディアの特徴量と各タスクの特徴量を表すパラメータとの第2関係を用いて、前記取得部で取得された前記パラメータから前記メディアにおける各タスクのタスク特徴量を予測する予測部と、を有することを特徴とする。

発明の効果

[0007] 本発明によれば、複数のタスクの推論を効率よく実行することが可能となる。

図面の簡単な説明

[0008] [図1]図1は、本実施形態の推論装置の概要を説明するための図である。
[図2]図2は、本実施形態の推論装置の概要を説明するための図である。
[図3]図3は、本実施形態の推論装置の概要を説明するための図である。
[図4]図4は、本実施形態の推論装置の概略構成を例示する模式図である。
[図5]図5は、推論装置の処理を説明するための図である。
[図6]図6は、推論装置の処理を説明するための図である。
[図7]図7は、推論処理手順を示すフローチャートである。
[図8]図8は、推論処理手順を示すフローチャートである。
[図9]図9は、推論処理手順を示すフローチャートである。
[図10]図10は、推論処理手順を示すフローチャートである。
[図11]図11は、推論処理手順を示すフローチャートである。
[図12]図12は、第2の実施形態を説明するための図である。
[図13]図13は、第2の実施形態の推論処理手順を示すフローチャートである。
[図14]図14は、実施例を説明するための図である。
[図15]図15は、推論プログラムを実行するコンピュータの一例を示す図で

ある。

発明を実施するための形態

[0009] 以下、図面を参照して、本発明の一実施形態を詳細に説明する。なお、この実施形態により本発明が限定されるものではない。また、図面の記載において、同一部分には同一の符号を付して示している。

[0010] [推論装置の概要]

図1は、本実施形態の推論装置の概要を説明するための図である。本実施形態において、メディアとは、例えば、画像、音声、自然言語のいずれかである。図1には、メディアに対する単一のタスクの予測に、TransformerとObject Queryとを用いる手法が例示されている。この手法における例えば画像に対するモデルは、画像分類用に作られたバックボーンと呼ばれるCNN (Convolutional Neural Network) 部分と、各タスクの予測を行うタスクヘッドと呼ばれる部分からなる。そして、バックボーンの入力である画像特徴量と、タスクに関するパラメータであるObject QueryとをTransformerに入力することにより、各タスクの出力を得る。

[0011] ここで、Transformerは、図1に例示するように、Transformer EncoderとTransformer Decoderとで構成される。Transformer Encoderは、複数のタスクのそれぞれのモデルに共通に設計されており、入力された特徴量間の類似度をもとに特徴量を抽出することにより、入力のうちの注目すべき箇所を学習する仕組みであるAttentionによる非線形変換で、バックボーンの入力である画像特徴量を精製する。

[0012] また、Transformer Decoderは、Transformer Encoderから出力された画像特徴量とObject Queryとの入力を受け付けて、複数層(L層)の処理を繰り返す。これにより、各Object Queryがそれぞれの予測の状態の把握と画像内での役割の認識とを行いつつ、徐々に所望のタスクの予測になっていく。

- [0013] このTransformer Decoderは、主にSelf-AttentionとCross-Attentionとで構成される。Self-Attentionでは、複数のObject Queryのそれぞれに対応する予測の間の関係を抽出する。また、Cross-Attentionでは、画像特徴量と複数のObject Queryのそれぞれに対応する予測との間の関係を抽出する。ここで、Attentionは、入力された特徴量間の類似度をもとに特徴量を抽出することにより、非線形変換を行う機構である。
- [0014] Self-Attentionの処理は、各タスクのモデルでほとんど同じであることが多く、Cross-Attentionの処理は、タスクや手法によって異なることが多い。そのため、一般にTransformerを用いた手法は、様々なタスクで有効である一方、1つのタスクに特化したモデルを作成する必要がある。
- [0015] そこで、本実施形態に係る推論装置は、タスクの数だけタスクに特化したモデルを作成して推論を行う必要がありタスクの数に応じたコストがかかっていた従来のTransformerとObject Queryとを用いる手法に対して特定の改善を提供するものであり、マルチタスクAIに係る技術分野の向上を示すものである。
- [0016] ここで、図2および図3には、本実施形態の推論装置における処理が例示されている。具体的には、図2に例示するように、推論装置では、複数のタスクを同時に解くために、Transformer DecoderとObject Queryとがマルチタスクに拡張されている。
- [0017] すなわち、推論装置では、図3(b)に例示するように、各タスクに対応して、タスクの数と同数の種類のObject Queryを用意して、Transformer DecoderのSelf-Attentionにまとめて入力し、Object Queryの予測間の関係を抽出する処理を行う。これにより、複数のタスクの予測間の関係を抽出可能となる。
- [0018] また、Cross-Attention(CA)では、タスクの数だけ分

岐させ、各タスクに対応した種類のObject Queryごとに、画像特徴量とObject Queryの予測との間の関係を抽出する処理を行う。これにより、各タスクについて独自に発展してきたCross-Attentionの技術を取り込むことが可能となる。このように、推論装置では、マルチタスクのObject Queryを用いて、Transformer Decoderにおけるタスクを横断したSelf-Attentionの処理と、分岐したタスク独自のCross-Attentionの処理を行う。これにより、図3(a)に例示した従来の1つのタスクに対する処理と比較して、効率的にマルチタスクの推論を行うことが可能となる。

[0019] [推論装置の構成]

図4は、本実施形態の推論装置の概略構成を例示する模式図である。また、図5および図6は、推論装置の処理を説明するための図である。まず、図4に例示するように、本実施形態の推論装置10は、パソコン等の汎用コンピュータで実現され、入力部11、出力部12、通信制御部13、記憶部14、および制御部15を備える。

[0020] 入力部11は、キーボードやマウス等の入力デバイスを用いて実現され、操作者による入力操作に対応して、制御部15に対して処理開始などの各種指示情報を入力する。出力部12は、液晶ディスプレイなどの表示装置、プリンター等の印刷装置等によって実現される。例えば、出力部12には、後述する推論処理の結果が表示される。

[0021] 通信制御部13は、NIC (Network Interface Card) 等で実現され、LAN (Local Area Network) やインターネットなどの電気通信回線を介した外部の装置と制御部15との通信を制御する。例えば、通信制御部13は、各種情報を管理する管理装置等と制御部15との通信を制御する。

[0022] 記憶部14は、RAM (Random Access Memory)、フラッシュメモリ (Flash Memory) 等の半導体メモリ素子、または、ハードディスク、光ディスク等の記憶装置によって実現される。記憶部14には、推論装置10を動作させる処理プログラムや、処理プログラムの実行中に使用されるデータなど

が予め記憶され、あるいは処理の都度一時的に記憶される。なお、記憶部 14 は、通信制御部 13 を介して制御部 15 と通信する構成でもよい。

[0023] 本実施形態において、記憶部 14 は、後述する推論処理に用いられる、画像 14 a、画像特徴生成用パラメータ 14 b、画像特徴精製用パラメータ 14 c、予測生成用パラメータ 14 d、クエリ用パラメータ 14 e 等を記憶する。これらの情報は、後述する推論処理に先立って予め、あるいは推論処理の最中に、入力部 11 を介して、あるいは通信制御部 13 を介して各種情報を管理する管理装置等から取得され、記憶部 14 に格納される。

[0024] 画像 14 a には、後述する推論処理に入力される処理対象の画像が格納される。

[0025] 画像特徴生成用パラメータ 14 b には、後述する推論処理に用いられる画像特徴生成用パラメータが格納される。画像特徴生成用パラメータは、ニューラルネットワークモデルのパラメータであり、タスクの学習用データを用いた一般的なニューラルネットワークの教師あり学習によって得られる。

[0026] 画像特徴精製用パラメータ 14 c には、後述する推論処理に用いられる画像特徴精製用パラメータが格納される。画像特徴精製用パラメータは、ニューラルネットワークモデルのパラメータであり、画像特徴生成用パラメータと同様に、タスクの学習用データを用いた一般的なニューラルネットワークの教師あり学習によって得られる。

[0027] 予測生成用パラメータ 14 d には、後述する推論処理に用いられる予測生成用パラメータが格納される。予測生成用パラメータは、ニューラルネットワークモデルのパラメータであり、画像特徴生成用パラメータと同様に、タスクの学習用データを用いた一般的なニューラルネットワークの教師あり学習によって得られる。

[0028] クエリ用パラメータ 14 e には、後述する推論処理に用いられるクエリ用パラメータが格納される。クエリ用パラメータは、例えば `Object Query` 等のタスクの特徴量を表すパラメータであり、複数の各タスクに対応してタスク数 N と同数種類のクエリ用パラメータが格納されている。図 5

に示す例では、タスク1クエリ用パラメータ、タスク2クエリ用パラメータ、…、タスクNクエリ用パラメータのN種類のクエリ用パラメータが格納されている。

[0029] 各タスクに対応したクエリ用パラメータは、ランダムに初期化され、ニューラルネットワークの一般的な教師あり学習を経て得られる。また、クエリ用パラメータは、後述する推論処理において処理されるベクトルと同じ次元のベクトルである。

[0030] 制御部15は、CPU (Central Processing Unit) やNP (Network Processor) やFPGA (Field Programmable Gate Array) 等を用いて実現され、メモリに記憶された処理プログラムを実行する。これにより、制御部15は、図4に例示するように、取得部15a、画像特徴生成部15b、画像特徴精製部15c、予測生成部15d、予測間関係抽出部15eおよび画像-予測間関係抽出部15fとして機能して、推論処理を実行する。なお、これらの機能部は、それぞれ、あるいは一部が異なるハードウェアに実装されてもよい。また、制御部15は、その他の機能部を備えてもよい。

[0031] 取得部15aは、入力部11あるいは通信制御部13を介して、後述する推論処理に用いられる処理対象の画像、画像特徴生成用パラメータ、画像特徴精製用パラメータ、予測生成用パラメータ、クエリ用パラメータ等を記憶部14から取得する。

[0032] なお、取得部15aに代えて、後述する機能部が処理を行う際に上記のそれぞれの情報を記憶部14から取得してもよい。

[0033] 画像特徴生成部15bは、図5に例示するように、記憶部14から取得された画像特徴生成用パラメータと、処理対象の画像とを用いて、処理対象の画像に対して画像特徴生成用パラメータに基づいて画像特徴量を生成する。画像特徴生成部15bは、例えば、Transformer Encoderに相当する。

[0034] ここで、処理対象の画像は、縦×横×RGBの3次元ベクトルとして取得される。また、画像特徴生成部15bは、複数の画像に対する処理を行う場

合には、画像特徴生成用パラメータは、最初の1回だけ取得すればよい。

[0035] 画像特徴精製部15cは、図5に例示するように、記憶部14から取得された画像特徴精製用パラメータに基づいて、画像特徴生成部15bが生成した画像特徴量をさらに精製する。具体的には、画像特徴精製部15cは、Transformer Encoderに相当し、メディアの特徴量間の類似度を用いてさらに特徴量を抽出する機構であるAttentionに基づいて非線形変換を行うことにより、該メディアの特徴量を精製する。

[0036] ここで、画像特徴精製部15cは、複数の画像の画像特徴量に対する処理を行う場合には、画像特徴精製用パラメータは、最初の1回だけ取得すればよい。

[0037] なお、前述の取得部15aが、精製された画像特徴量を取得して、以下に説明する予測生成部15dに転送してもよい。この場合に、取得部15aが、例えば画像等の所定のメディアに対する複数のタスクについて、メディアの特徴量と、各タスクの特徴量を表すタスク用パラメータとを取得する。

[0038] 予測生成部15dは、異なるタスクを横断した各タスクの特徴量を表すクエリ用パラメータの間の第1関係、およびメディアの特徴量と各タスクの特徴量を表すクエリ用パラメータとの第2関係を用いて、取得部15aで取得されたクエリ用パラメータからメディアにおける各タスクのタスク特徴量を予測する。

[0039] 具体的には、予測生成部15dは、図5に例示するように、記憶部14から取得された予測生成用パラメータとクエリ用パラメータと、画像特徴精製部15cが精製した画像特徴量とを用いて、所望のタスクの予測を行う。

[0040] ここで、予測生成部15dは、複数の画像の画像特徴量に対する処理を行う場合には、予測生成用パラメータおよびクエリ用パラメータは、最初の1回だけ取得すればよい。

[0041] また、クエリ用パラメータは、図5に例示するように、複数の各タスクに対応したクエリ用パラメータが存在する。例えば、物体検出、姿勢推定の2つのタスクを推論処理の対象とする場合に、物体検出のクエリ用パラメータ

と、姿勢推定のクエリ用パラメータとが、予測生成部15dに入力され、画像特徴精製部15cが精製した画像特徴量から、物体検出、姿勢推定の2つのタスクの予測結果が出力される。

[0042] 予測生成部15dは、第1関係としてクエリ用パラメータの間の類似度を用いてパラメータの変換を行い、第2関係としてメディアの特徴量とクエリ用パラメータとの間の類似度を用いてパラメータの変換を行うことを繰り返すことで、タスク特徴量を予測する。

[0043] すなわち、予測生成部15dは、各タスクのクエリ用パラメータの間の類似度を用いてさらに特徴量を抽出するAttention機構に基づいて各タスクのクエリ用パラメータの非線形変換を行い、メディアの特徴量と各タスクのクエリ用パラメータとの間の類似度を用いてさらに特徴量を抽出するAttention機構に基づいて各タスクのクエリ用パラメータの非線形変換を行う。予測生成部15dは、この処理を繰り返すことにより、メディアにおける各タスクのタスク特徴量を予測する。

[0044] 具体的には、予測生成部15dは、あらかじめ学習済みのニューラルネットワークであるTransformerのDecoderであり、第1関係をSelf-Attention機構により抽出し、第2関係をCross-Attention機構により抽出する。

[0045] すなわち、予測生成部15dは、TransformerのDecoderであり、Self-Attentionが異なるタスクを横断して各タスクのクエリ用パラメータの間の関係を抽出し、Cross-Attentionがタスクごとにメディアの特徴量と各タスクのクエリ用パラメータとの関係を抽出することにより、メディアにおける各タスクのタスク特徴量を予測する。

[0046] ここで、図6は、予測生成部の処理を説明するための図である。図6に例示するように、予測生成部15dは、予測間関係抽出部15eと画像-予測間関係抽出部15fとを含む。予測間関係抽出部15eは、例えば、Transformer DecoderのSelf-Attentionに相当

し、タスクの種類や数によらず横断的に同じ処理を行う。一方、画像－予測間関係抽出部15fは、例えば、Transformer DecoderのCross-Attentionに相当し、タスクの種類によって異なる処理を行うため、タスクの数と同数の画像－予測間関係抽出部15fが並列に処理を行う。

[0047] 予測生成部15dには、1対の予測間関係抽出部15eと画像－予測間関係抽出部15fとのペアからなる層が所定数Lだけ存在し、各層を経るごとにクエリ用パラメータが徐々に変化し、最終的に各タスクの推論結果である予測が出力される。

[0048] 具体的には、まず、1層目において、予測間関係抽出部15eは、予測生成用パラメータに基づいて、各クエリ用パラメータ間の関係を抽出し、抽出した関係をもとに各クエリ用パラメータを更新する（以下、更新されたクエリ用パラメータを「クエリ用パラメータ（予測）」と記す場合もある）。次に、画像－予測間関係抽出部15fが、予測生成用パラメータに基づいて、画像特徴量と各タスクの更新されたクエリ用パラメータ（予測）のとの間の関係を抽出し、抽出した関係をもとに各タスクのクエリ用パラメータを更新する。

[0049] 次に、2層目において、予測間関係抽出部15eは、更新された予測生成用パラメータに基づいて、各クエリ用パラメータ（予測）の間の関係を抽出し、抽出した関係をもとに各クエリ用パラメータを更新する。次に、画像－予測間関係抽出部15fが、予測生成用パラメータに基づいて、画像特徴量と各タスクの更新されたクエリ用パラメータ（予測）のとの間の関係を抽出し、抽出した関係をもとに各タスクのクエリ用パラメータを更新する。

[0050] このように、予測間関係抽出部15eおよび画像－予測間関係抽出部15fは、同様の処理をL層まで繰り返すことにより、各タスクの推論結果である予測を出力する。このように、推論装置10によれば、メディアに対する複数のタスクの推論を効率よく実行可能となる。また、各種パラメータの学習も、複数のタスクについてまとめて行えるので、効率よく実行可能となる。

。

[0051] [推論処理]

次に、図 7～図 11 を参照して、本実施形態に係る推論装置 10 による推論処理について説明する。図 7～図 11 は、推論処理手順を示すフローチャートである。まず、図 7 には、図 7 のフローチャートは、例えば、ユーザが開始を指示する操作入力を行ったタイミングで開始される。

[0052] まず、取得部 15 a が、推論処理に用いられる処理対象の画像、画像特徴生成用パラメータ、画像特徴精製用パラメータ、予測生成用パラメータ、クエリ用パラメータ等を記憶部 14 から取得する。

[0053] そして、画像特徴生成部 15 b が、画像特徴生成用パラメータと、処理対象の画像とを用いて、処理対象の画像に対して画像特徴生成用パラメータに基づいて画像特徴量を生成する（ステップ S 1）。

[0054] また、画像特徴精製部 15 c が、画像特徴精製用パラメータに基づいて、画像特徴生成部 15 b が生成した画像特徴量をさらに精製する（ステップ S 2）。

[0055] 次に、予測生成部 15 d が、予測生成用パラメータを用いて、精製された画像特徴量と各タスクのクエリ用パラメータとを用いて、画像における各タスクの特徴量の予測を生成し（ステップ S 3）、生成した予測を出力する（ステップ S 4）。これにより、一連の推論処理が終了する。

[0056] 次に、図 8 には、図 7 に示したステップ S 1 の画像特徴生成部の処理手順が例示されている。まず、画像特徴生成部 15 b は、取得部 15 a あるいは記憶部 14 から、画像特徴生成用パラメータと処理対象の入力画像とを取得する（ステップ S 101～S 102）。

[0057] 次に、画像特徴生成部 15 b は、処理対象の入力画像に対して画像特徴生成用パラメータに基づいて画像特徴量を生成する（ステップ S 103）。また、画像特徴生成部 15 b は、生成した画像特徴量を後段の画像特徴精製部 15 c に対して出力して（ステップ S 104）、一連の処理を終了する。

[0058] 次に、図 9 には、図 7 に示したステップ S 2 の画像特徴生成部の処理手順

が例示されている。まず、画像特徴精製部 15 c は、取得部 15 a あるいは記憶部 14 から、画像特徴精製用パラメータを取得し（ステップ S 201）、画像特徴生成部 15 b から生成された画像特徴量を取得する（ステップ S 202）。

[0059] 次に、画像特徴精製部 15 c は、画像特徴量に対して画像特徴精製用パラメータに基づいて画像特徴量を精製する（ステップ S 203）。また、画像特徴生成部 15 b は、精製した画像特徴量を後段の予測生成部 15 d に対して出力して（ステップ S 204）、一連の処理を終了する。

[0060] 次に、図 10 には、図 7 に示したステップ S 3～S 4 の予測生成部の処理手順が例示されている。まず、予測生成部 15 d は、取得部 15 a あるいは記憶部 14 から、予測生成用パラメータと、クエリ用パラメータを取得し（ステップ S 301～S 302）、画像特徴精製部 15 c から精製された画像特徴量を取得する（ステップ S 303）。

[0061] 次に、予測生成部 15 d は、画像特徴量とクエリ用パラメータとに対し、予測生成用パラメータに基づいてタスクの特徴量の予測を生成する（ステップ S 304）。そして、予測生成部 15 d は、生成したタスクの予測を出力して（ステップ S 305）、一連の処理を終了する。

[0062] また、図 11 には、図 10 に示したステップ S 304 の処理の詳細な処理手順が例示されている。ステップ S 303 の処理で画像特徴精製部 15 c から精製された画像特徴量を取得した後、予測生成部 15 d は、図 6 に示した 1 層目の処理を行う。具体的には、予測間関係抽出部 15 e が、各タスクのクエリ用パラメータに対して、予測生成用パラメータに基づいてクエリ用パラメータ間の関係を抽出し、抽出した関係をもとに各クエリ用パラメータを更新する（ステップ S 401）。次に、画像－予測間関係抽出部 15 f が、画像特徴量と更新された各タスクのクエリ用パラメータとの間の関係を抽出し、抽出した関係をもとにさらに各タスクのクエリ用パラメータを更新する（ステップ S 402）。

[0063] 次に、予測生成部 15 d は、2 層目の処理を行う。すなわち、予測間関係

抽出部15eが、更新された各タスクのクエリ用パラメータに対して、予測生成用パラメータに基づいてクエリ用パラメータ間の関係を抽出し、抽出した関係をもとに各クエリ用パラメータを更新する。また、画像ー予測間関係抽出部15fが、画像特徴量と更新された各タスクのクエリ用パラメータとの間の関係を抽出し、抽出した関係をもとにさらに各タスクのクエリ用パラメータを更新する。

[0064] このように、予測生成部15dは、L層目まで同様の処理を繰り返し（ステップS403）、L層目で更新されたクエリ用パラメータを生成したタスクの予測とする（ステップS404）。その後、予測生成部15dは、生成したタスクの予測を出力して（ステップS305）、一連の処理を終了する。

[0065] なお、上記の処理におけるメディアは、画像に限定されない。上記のとおり、メディアとして、音声、自然言語にも同様の処理を適用可能である。その場合には、各メディアの処理を行う際に一般的な特徴量抽出を行って、モデルへの入力とすればよい。

[0066] [第2の実施形態]

図12は、第2の実施形態の推論装置の処理を説明するための図である。なお以下では、上記の第1の実施形態の推論装置10の推論処理と異なる点についてのみ説明を行い、共通する点についての説明を省略する。

[0067] 第2の本実施形態に係る推論装置10も、上記実施形態の推論装置10と同様に、タスクの数だけタスクに特化したモデルを作成して推論を行う必要がありタスクの数に応じたコストがかかっていた従来のTransformerとObject Queryとを用いる手法に対して特定の改善を提供するものであり、マルチタスクAIに係る技術分野の向上を示すものである。

[0068] 具体的には、上記実施形態の推論装置10では、図6に示したように、予測生成部15dでは、予測間関係抽出部15eが処理を行った後に、画像ー予測間関係抽出部15fが処理を行っているが、この処理順に限定されない

。すなわち、図 1 2 に示すように、画像－予測間関係抽出部 1 5 f が処理を行った後に、予測間関係抽出部 1 5 e が処理を行ってもよい。

[0069] この場合にも、上記実施形態の推論装置 1 0 と同様に、メディアに対する複数のタスクの推論を効率よく実行可能となる。また、各種パラメータの学習も、複数のタスクについてまとめて行えるので、効率よく実行可能となる。

[0070] 図 1 3 は、第 2 の実施形態の推論処理手順を示すフローチャートとである。図 1 3 に示すように、図 1 1 のフローチャートとは、ステップ S 4 1 1 ～ S 4 1 3 の処理が異なる。

[0071] 具体的には、ステップ S 3 0 3 の処理で画像特徴精製部 1 5 c から精製された画像特徴量を取得した後、予測生成部 1 5 d は、図 1 2 に示した 1 層目の処理を行う。具体的には、画像－予測間関係抽出部 1 5 f が、画像特徴量と各タスクのクエリ用パラメータとの間の関係を抽出し、抽出した関係をもとに各クエリ用パラメータを更新する（ステップ S 4 1 1）。次に、予測間関係抽出部 1 5 e が、更新された各タスクのクエリ用パラメータに対して、予測生成用パラメータに基づいてクエリ用パラメータ間の関係を抽出し、抽出した関係をもとに各クエリ用パラメータを更新する（ステップ S 4 1 2）。

[0072] 次に、予測生成部 1 5 d は、2 層目の処理を行う。すなわち、画像－予測間関係抽出部 1 5 f が、画像特徴量と更新された各タスクのクエリ用パラメータとの間の関係を抽出し、抽出した関係をもとに各タスクのクエリ用パラメータを更新する。また、予測間関係抽出部 1 5 e が、更新された各タスクのクエリ用パラメータに対して、予測生成用パラメータに基づいてクエリ用パラメータ間の関係を抽出し、抽出した関係をもとに各クエリ用パラメータを更新する。

[0073] このように、予測生成部 1 5 d は、L 層目まで同様の処理を繰り返し（ステップ S 4 1 3）、L 層目で更新されたクエリ用パラメータを生成したタスクの予測とする（ステップ S 4 0 4）。その後、予測生成部 1 5 d は、生成

したタスクの予測を出力して（ステップS305）、一連の処理を終了する。

[0074] [効果]

以上、説明したように、本実施形態の推論装置10において、取得部15aが、所定のメディアに対する複数のタスクについて、メディアの特徴量と、各タスクの特徴量を表すクエリ用パラメータとを取得する。予測生成部15dは、異なるタスクを横断した各タスクの特徴量を表すクエリ用パラメータの間の第1関係、およびメディアの特徴量と各タスクの特徴量を表すクエリ用パラメータとの第2関係を用いて、取得部15aで取得されたクエリ用パラメータからメディアにおける各タスクのタスク特徴量を予測する。

[0075] 具体的には、予測生成部15dは、第1関係としてクエリ用パラメータの間の類似度を用いてパラメータの変換を行い、第2関係としてメディアの特徴量とクエリ用パラメータとの間の類似度を用いてパラメータの変換を行うことを繰り返すことで、タスク特徴量を予測する。

[0076] 例えば、予測生成部15dは、あらかじめ学習済みのニューラルネットワークであるTransformerのDecoderであり、第1関係をSelf-Attention機構により抽出し、第2関係をCross-Attention機構により抽出する。

[0077] これにより、複数のタスクの推論を効率よく実行可能となる。また、複数のタスクに対する各種パラメータの学習も、まとめて行えるので、効率よく実行可能となる。

[0078] また、メディアは、画像、音声、自然言語のいずれかである。このように、推論装置10によれば、多様なメディアにおける複数のタスクの推論処理が効率よく実行可能となる。

[0079] [実施例]

図14は、実施例を説明するための図である。本実施例では、タスクを物体検出および姿勢推定として、従来のTransformerを使ったモデルと上記実施形態の推論装置10との推論時間および学習時間を比較した。

従来のTransformerを使ったモデルとして、物体検出にはDeformable DETRを用い、姿勢推定にはPETR (Position Embedding TRansformation) を用いた。また、バックボーンとしてR-50を用いた。これに対し、上記実施形態のモデルのバックボーンとして、R-50を用いた場合の他、Swin-Tを用いた場合についても比較した。

[0080] 図14に例示するように、バックボーンが同じR-50である場合に、上記実施形態の推論装置10では、学習時間・推論時間ともに従来の70%に抑えられ、90%以上の精度を達成することが確認された。また、R-50より大きいバックボーンSwin-Tを使った場合には、推論装置10であ、従来の80%の学習時間・推論時間で、優れた精度を達成することが確認された。

[0081] [プログラム]

上記実施形態に係る推論装置10が実行する処理をコンピュータが実行可能な言語で記述したプログラムを作成することもできる。一実施形態として、推論装置10は、パッケージソフトウェアやオンラインソフトウェアとして上記の推論処理を実行する推論プログラムを所望のコンピュータにインストールさせることによって実装できる。例えば、上記の推論プログラムを情報処理装置に実行させることにより、情報処理装置を推論装置10として機能させることができる。ここで言う情報処理装置には、デスクトップ型またはノート型のパーソナルコンピュータが含まれる。また、その他にも、情報処理装置にはスマートフォン、携帯電話機やPHS (Personal Handyphone System) などの移動体通信端末、さらには、PDA (Personal Digital Assistant) などのスレート端末などがその範疇に含まれる。また、推論装置10の機能を、クラウドサーバに実装してもよい。

[0082] 図15は、推論プログラムを実行するコンピュータの一例を示す図である。コンピュータ1000は、例えば、メモリ1010と、CPU1020と、ハードディスクドライブインタフェース1030と、ディスクドライブインタフェース1040と、シリアルポートインタフェース1050と、ビデ

オアダプタ１０６０と、ネットワークインタフェース１０７０とを有する。
これらの各部は、バス１０８０によって接続される。

[0083] メモリ１０１０は、ＲＯＭ (Read Only Memory) １０１１およびＲＡＭ １０１２を含む。ＲＯＭ １０１１は、例えば、ＢＩＯＳ (Basic Input Output System) 等のブートプログラムを記憶する。ハードディスクドライブインタフェース１０３０は、ハードディスクドライブ１０３１に接続される。ディスクドライブインタフェース１０４０は、ディスクドライブ１０４１に接続される。ディスクドライブ１０４１には、例えば、磁気ディスクや光ディスク等の着脱可能な記憶媒体が挿入される。シリアルポートインタフェース１０５０には、例えば、マウス１０５１およびキーボード１０５２が接続される。ビデオアダプタ１０６０には、例えば、ディスプレイ１０６１が接続される。

[0084] ここで、ハードディスクドライブ１０３１は、例えば、ＯＳ (Operating System) １０９１、アプリケーションプログラム１０９２、プログラムモジュール１０９３およびプログラムデータ１０９４を記憶する。上記実施形態で説明した各情報は、例えばハードディスクドライブ１０３１やメモリ１０１０に記憶される。

[0085] また、推論プログラムは、例えば、コンピュータ１０００によって実行される指令が記述されたプログラムモジュール１０９３として、ハードディスクドライブ１０３１に記憶される。具体的には、上記実施形態で説明した推論装置１０が実行する各処理が記述されたプログラムモジュール１０９３が、ハードディスクドライブ１０３１に記憶される。

[0086] また、推論プログラムによる情報処理に用いられるデータは、プログラムデータ１０９４として、例えば、ハードディスクドライブ１０３１に記憶される。そして、ＣＰＵ １０２０が、ハードディスクドライブ１０３１に記憶されたプログラムモジュール１０９３やプログラムデータ１０９４を必要に応じてＲＡＭ １０１２に読み出して、上述した各手順を実行する。

[0087] なお、推論プログラムに係るプログラムモジュール１０９３やプログラム

データ1094は、ハードディスクドライブ1031に記憶される場合に限られず、例えば、着脱可能な記憶媒体に記憶されて、ディスクドライブ1041等を介してCPU1020によって読み出されてもよい。あるいは、推論プログラムに係るプログラムモジュール1093やプログラムデータ1094は、LANやWAN (Wide Area Network) 等のネットワークを介して接続された他のコンピュータに記憶され、ネットワークインタフェース1070を介してCPU1020によって読み出されてもよい。

[0088] 以上の実施形態に関し、更に以下の付記を開示する。

[0089] (付記項1)

メモリと、

前記メモリに接続された少なくとも1つのプロセッサと、

を含み、

前記プロセッサは、

所定のメディアに対する複数のタスクについて、前記メディアの特徴量と、各タスクの特徴量を表すパラメータとを取得し、

異なるタスクを横断した各タスクの特徴量を表すパラメータの間の第1関係、および前記メディアの特徴量と各タスクの特徴量を表すパラメータとの第2関係を用いて、取得された前記パラメータから前記メディアにおける各タスクのタスク特徴量を予測する。

推論装置。

[0090] (付記項2)

付記項1に記載の推論装置による推論処理を実行するようにコンピュータによって実行可能なプログラムを記憶した非一時的記憶媒体。

[0091] 以上、本発明者によってなされた発明を適用した実施形態について説明したが、本実施形態による本発明の開示の一部をなす記述および図面により本発明は限定されることはない。すなわち、本実施形態に基づいて当業者等によりなされる他の実施形態、実施例および運用技術等は全て本発明の範疇に含まれる。

符号の説明

- [0092] 1 0 推論装置
- 1 1 入力部
- 1 2 出力部
- 1 3 通信制御部
- 1 4 記憶部
- 1 4 a 画像
- 1 4 b 画像特徴生成用パラメータ
- 1 4 c 画像特徴精製用パラメータ
- 1 4 d 予測生成用パラメータ
- 1 4 e クエリ用パラメータ
- 1 5 制御部
- 1 5 a 取得部
- 1 5 b 画像特徴生成部
- 1 5 c 画像特徴精製部
- 1 5 d 予測生成部（予測部）
- 1 5 e 予測間関係抽出部
- 1 5 f 画像－予測間関係抽出部

請求の範囲

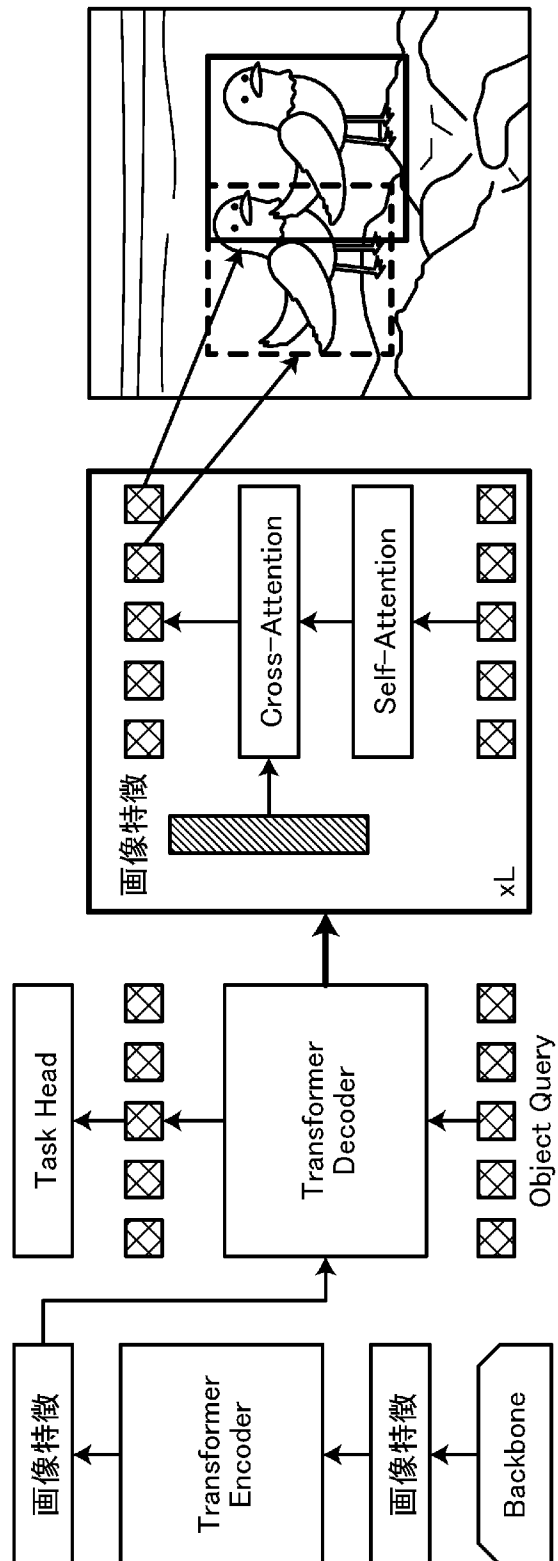
- [請求項1] 所定のメディアに対する複数のタスクについて、前記メディアの特徴量と、各タスクの特徴量を表すパラメータとを取得する取得部と、異なるタスクを横断した各タスクの特徴量を表すパラメータの間の第1関係、および前記メディアの特徴量と各タスクの特徴量を表すパラメータとの第2関係を用いて、前記取得部で取得された前記パラメータから前記メディアにおける各タスクのタスク特徴量を予測する予測部と、有することを特徴とする推論装置。
- [請求項2] 前記予測部は、前記第1関係として前記パラメータの間の類似度を用いて前記パラメータの変換を行い、前記第2関係として前記メディアの特徴量と前記パラメータとの間の類似度を用いて前記パラメータの変換を行うことを繰り返すことで、前記タスク特徴量を予測することを特徴とする請求項1に記載の推論装置。
- [請求項3] 前記予測部は、あらかじめ学習済みのニューラルネットワークであるTransformerのDecoderであり、前記第1関係をSelf-Attention機構により抽出し、前記第2関係をCross-Attention機構により抽出することを特徴とする請求項1に記載の推論装置。
- [請求項4] 前記メディアは、画像、音声、自然言語のいずれかであることを特徴とする請求項1に記載の推論装置。
- [請求項5] 推論装置が実行する推論方法であって、所定のメディアに対する複数のタスクについて、前記メディアの特徴量と、各タスクの特徴量を表すパラメータとを取得する取得工程と、異なるタスクを横断した各タスクの特徴量を表すパラメータの間の第1関係、および前記メディアの特徴量と各タスクの特徴量を表すパラメータとの第2関係を用いて、前記取得工程で取得された前記パラ

メータから前記メディアにおける各タスクのタスク特徴量を予測する
予測工程と、

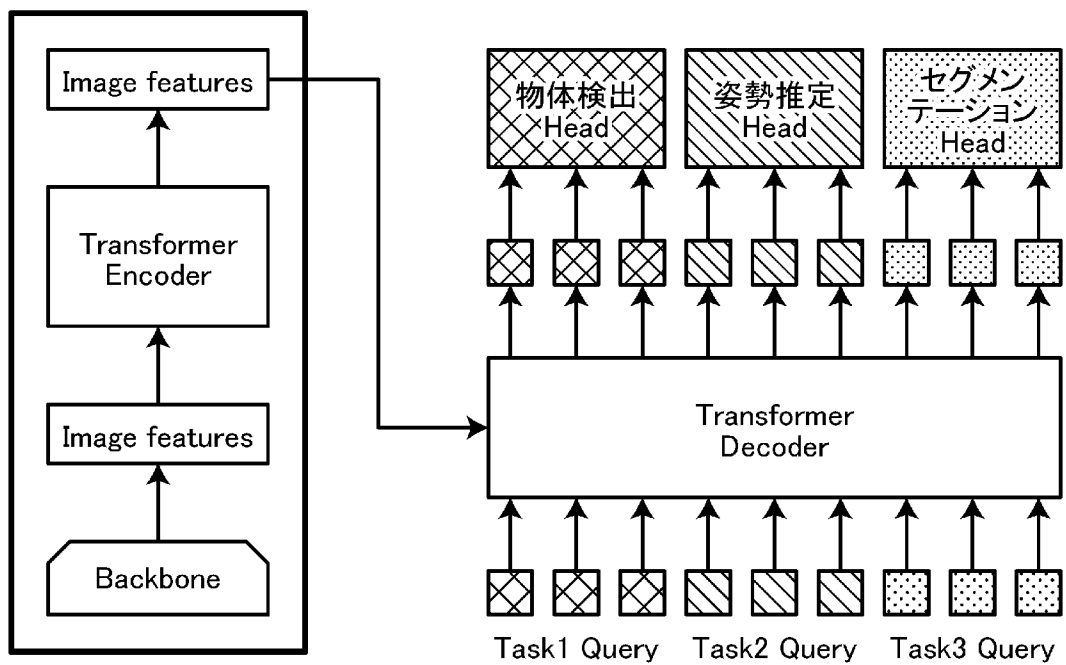
を含んだことを特徴とする推論方法。

[請求項6] コンピュータを請求項1～4のいずれか1項に記載の推論装置として機能させるための推論プログラム。

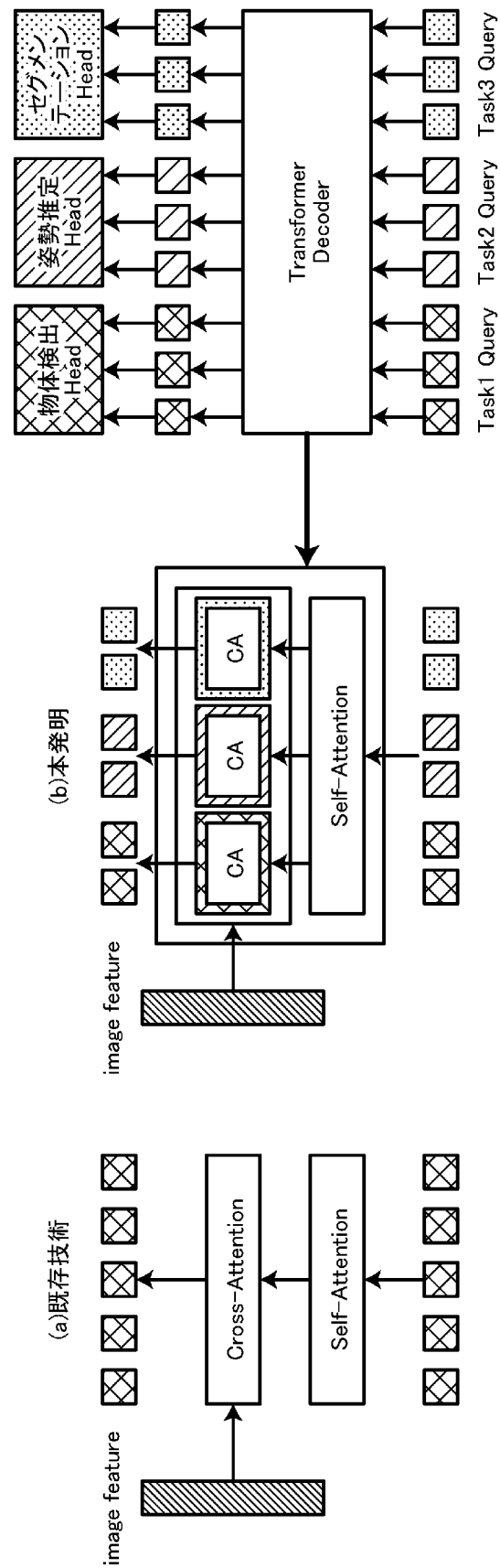
[図1]



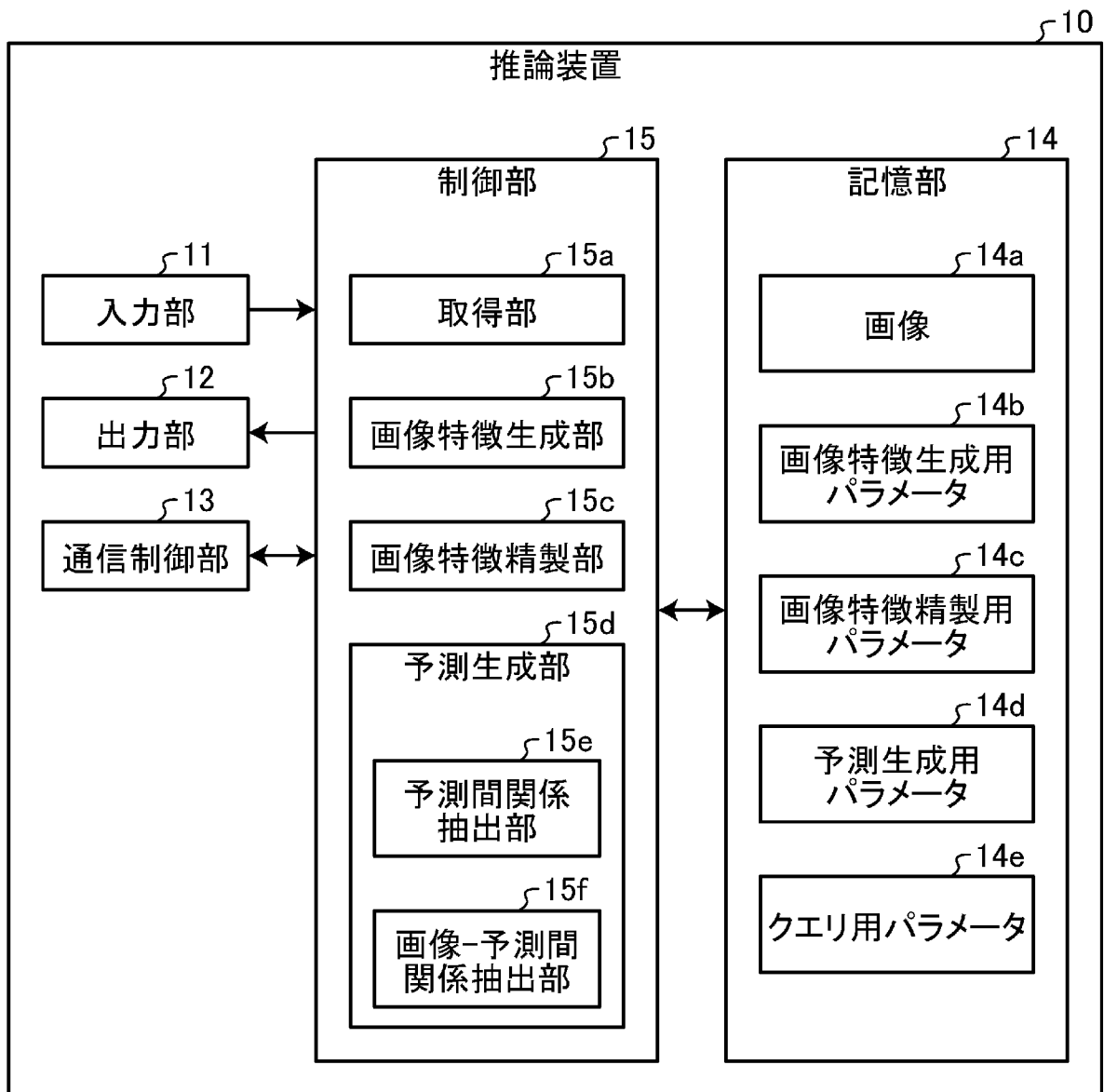
[図2]



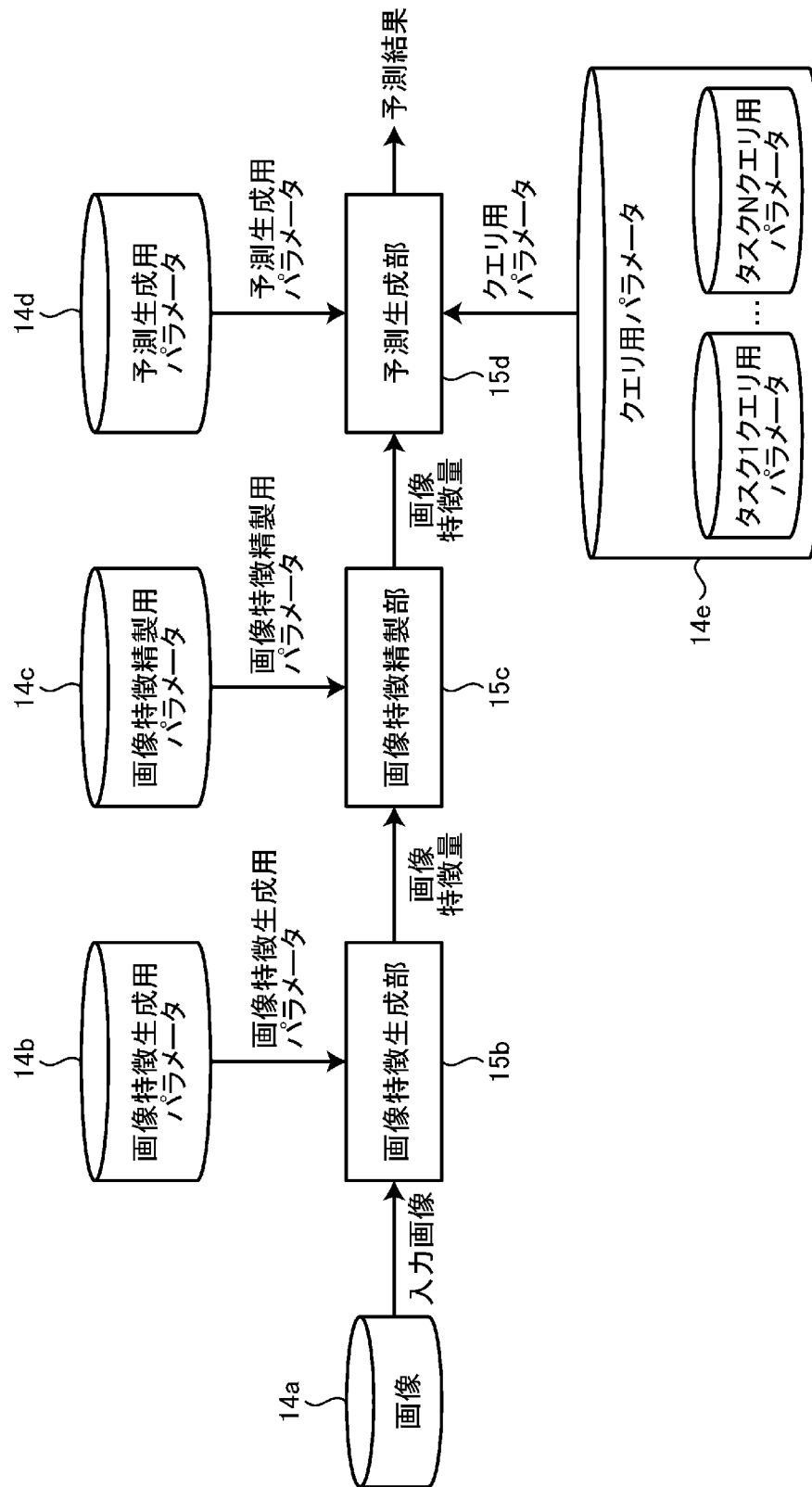
[図3]



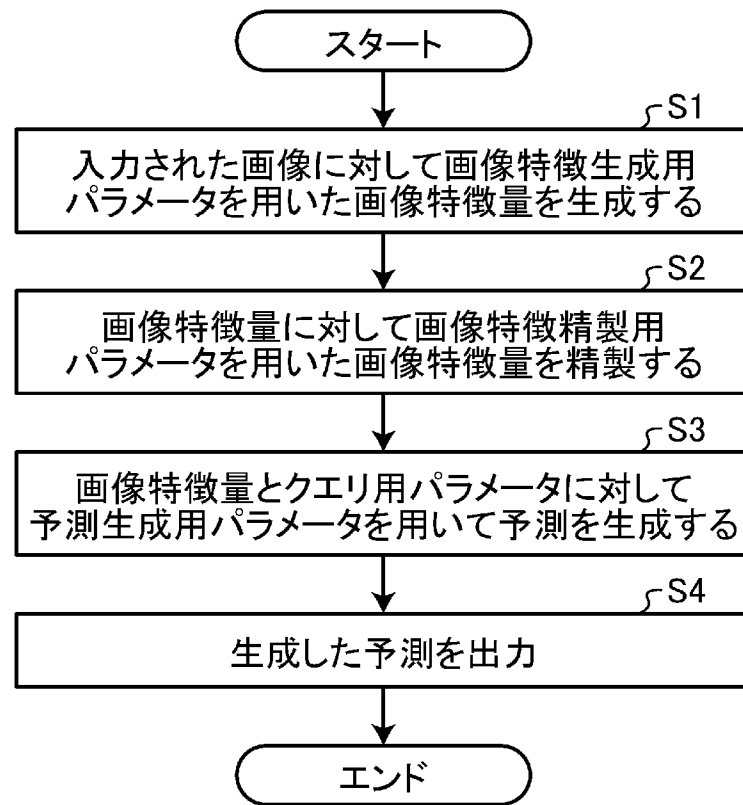
[図4]



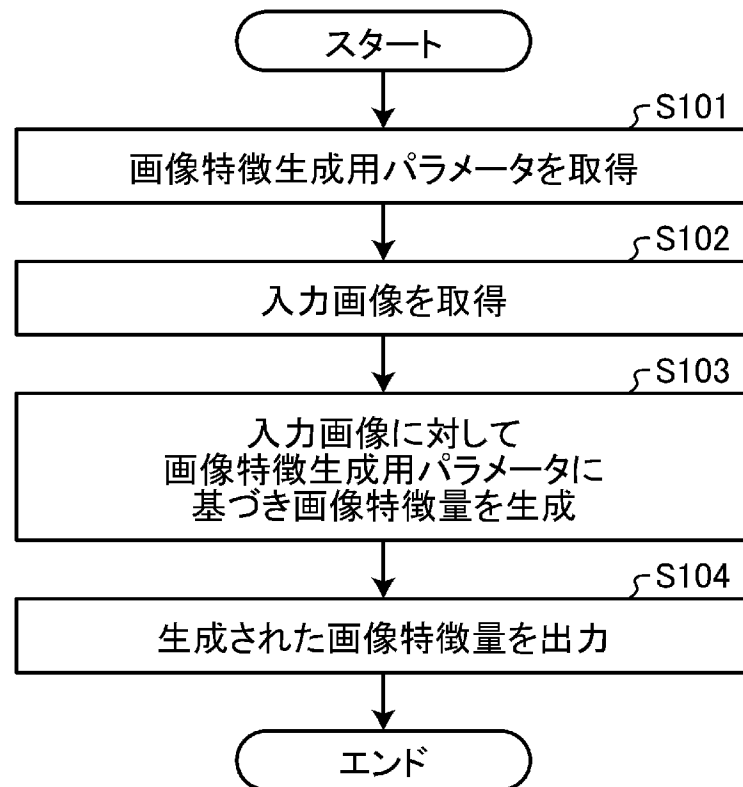
[図5]



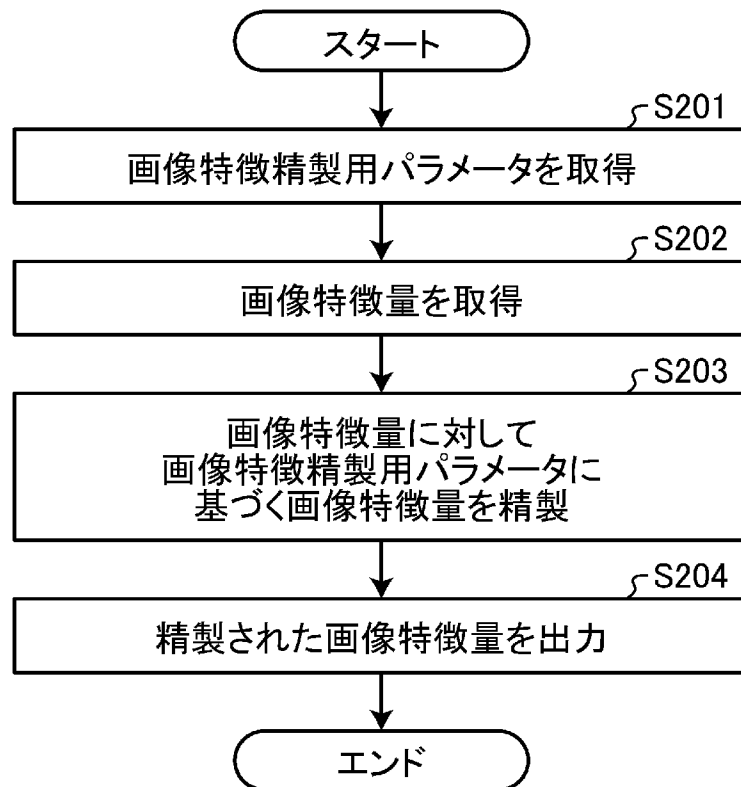
[図7]



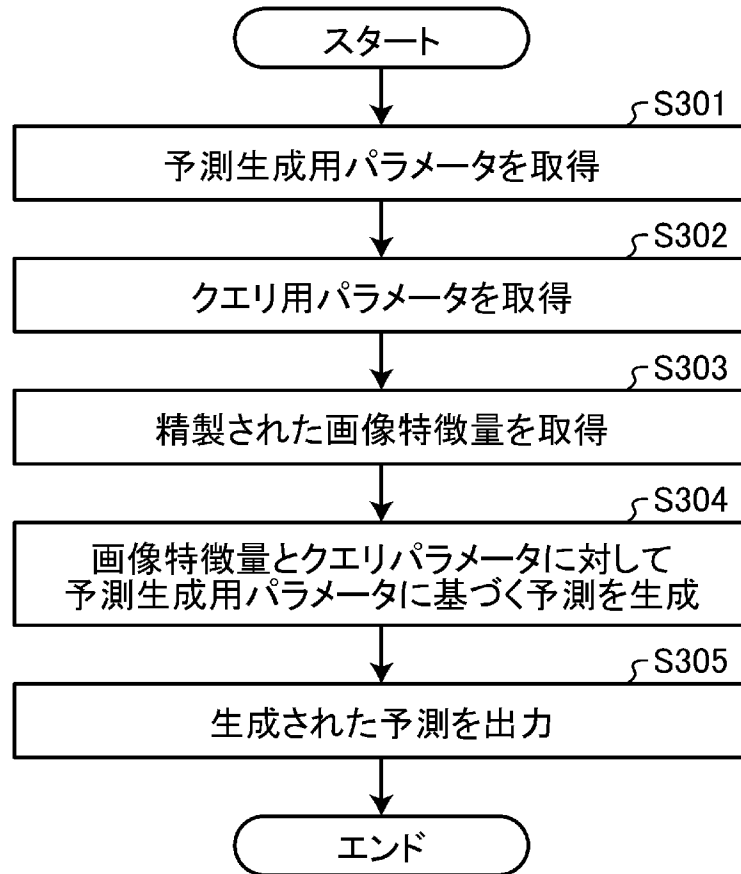
[図8]



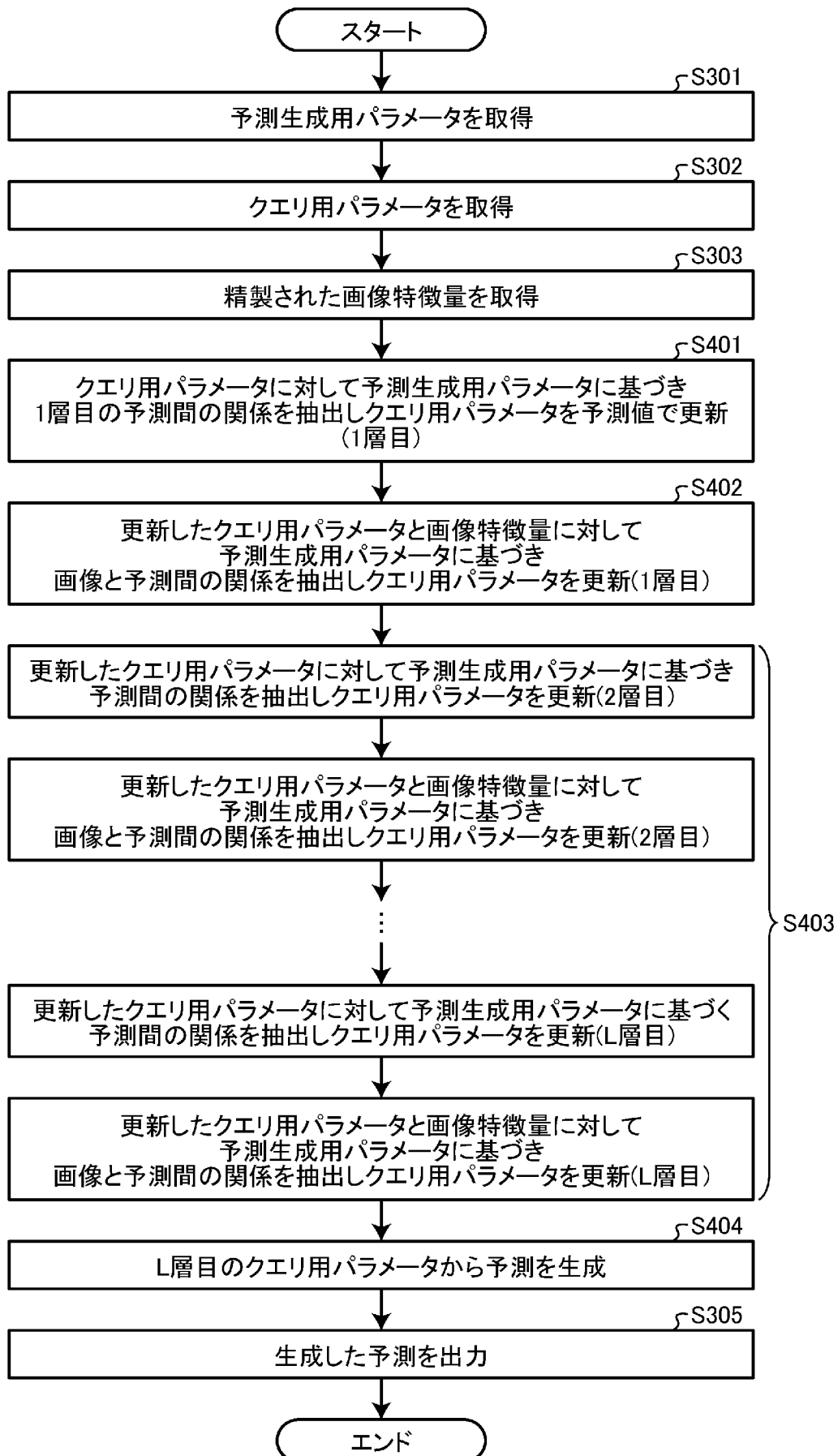
[図9]



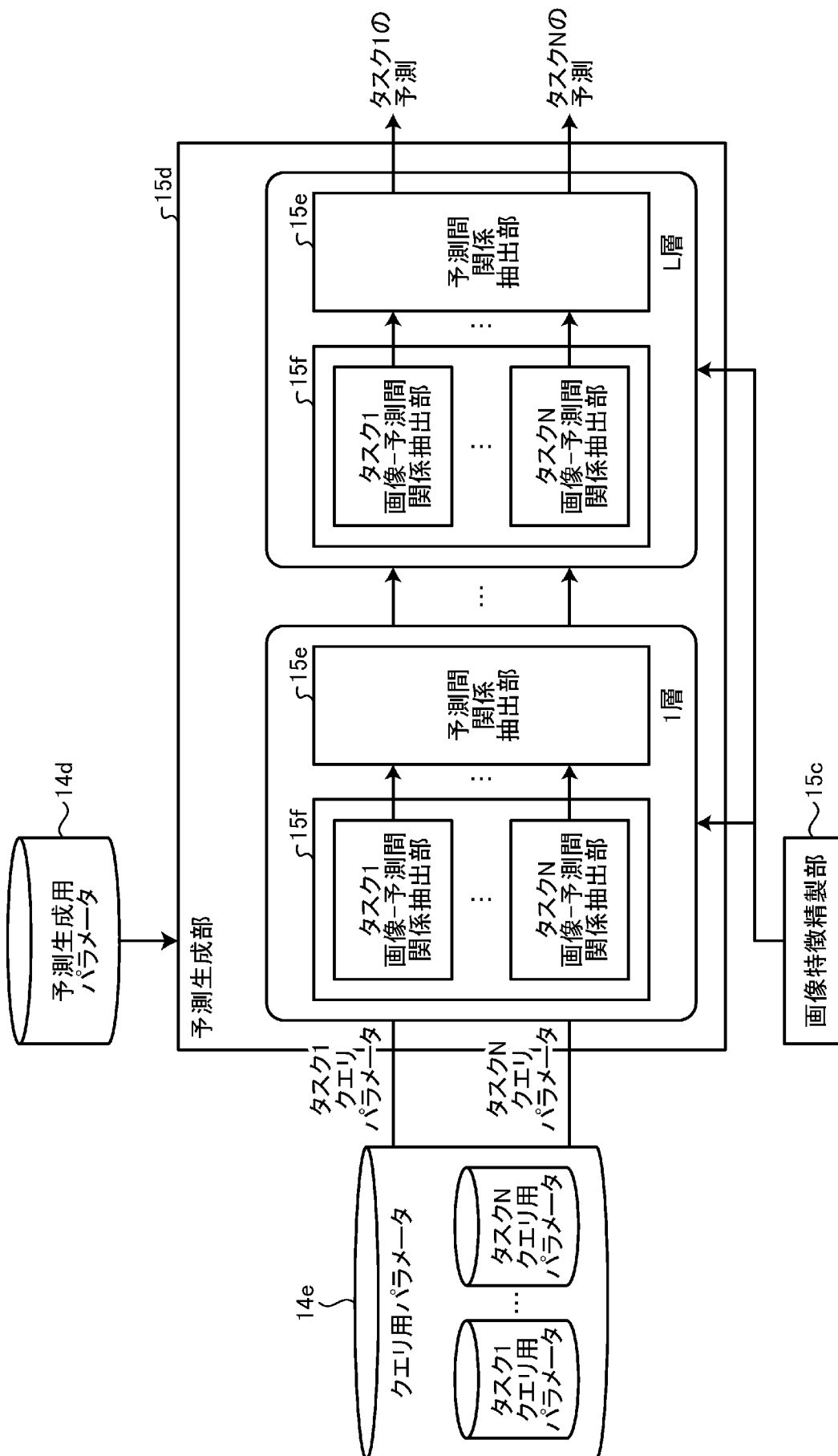
[図10]



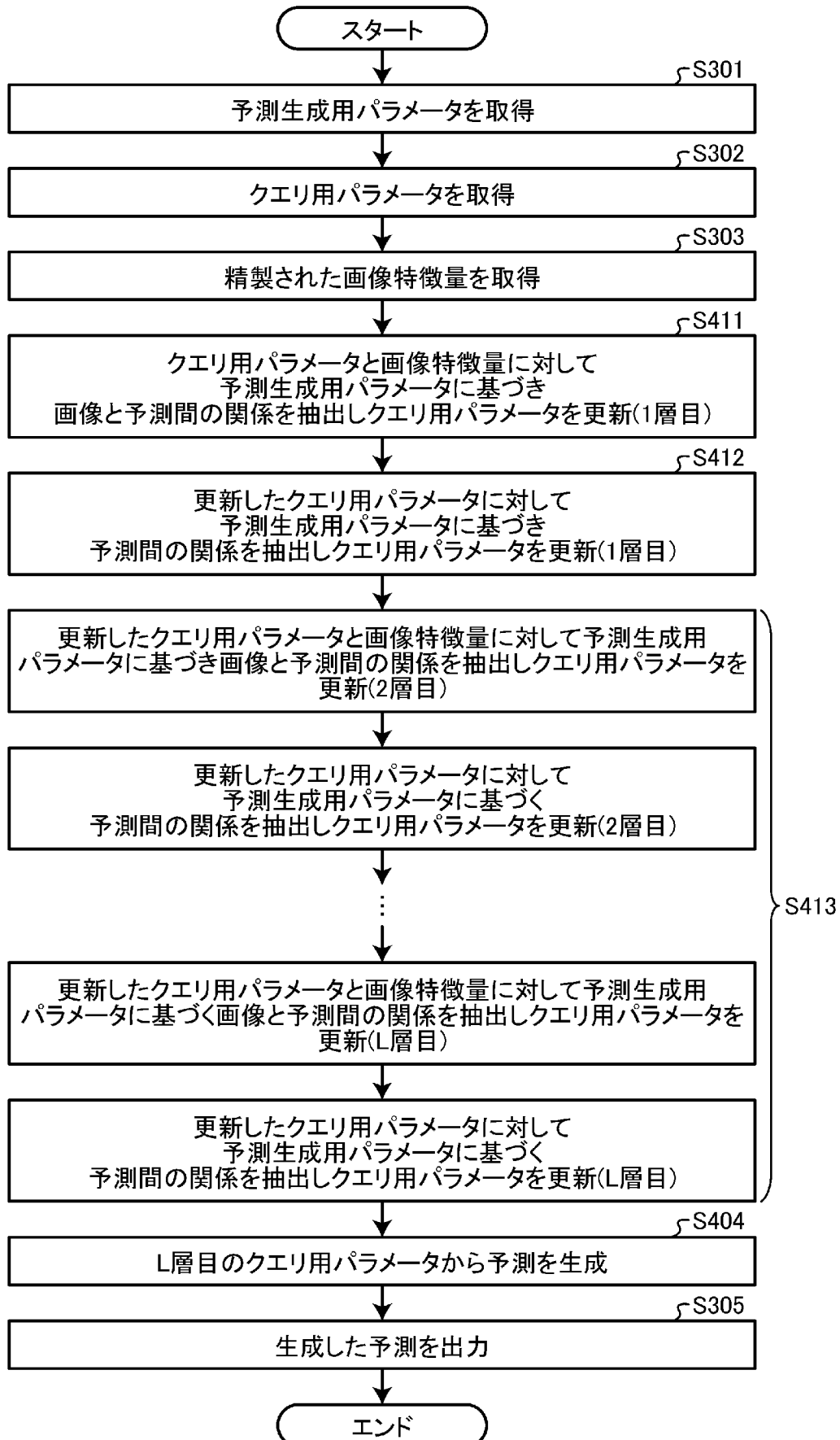
[図11]



[図12]



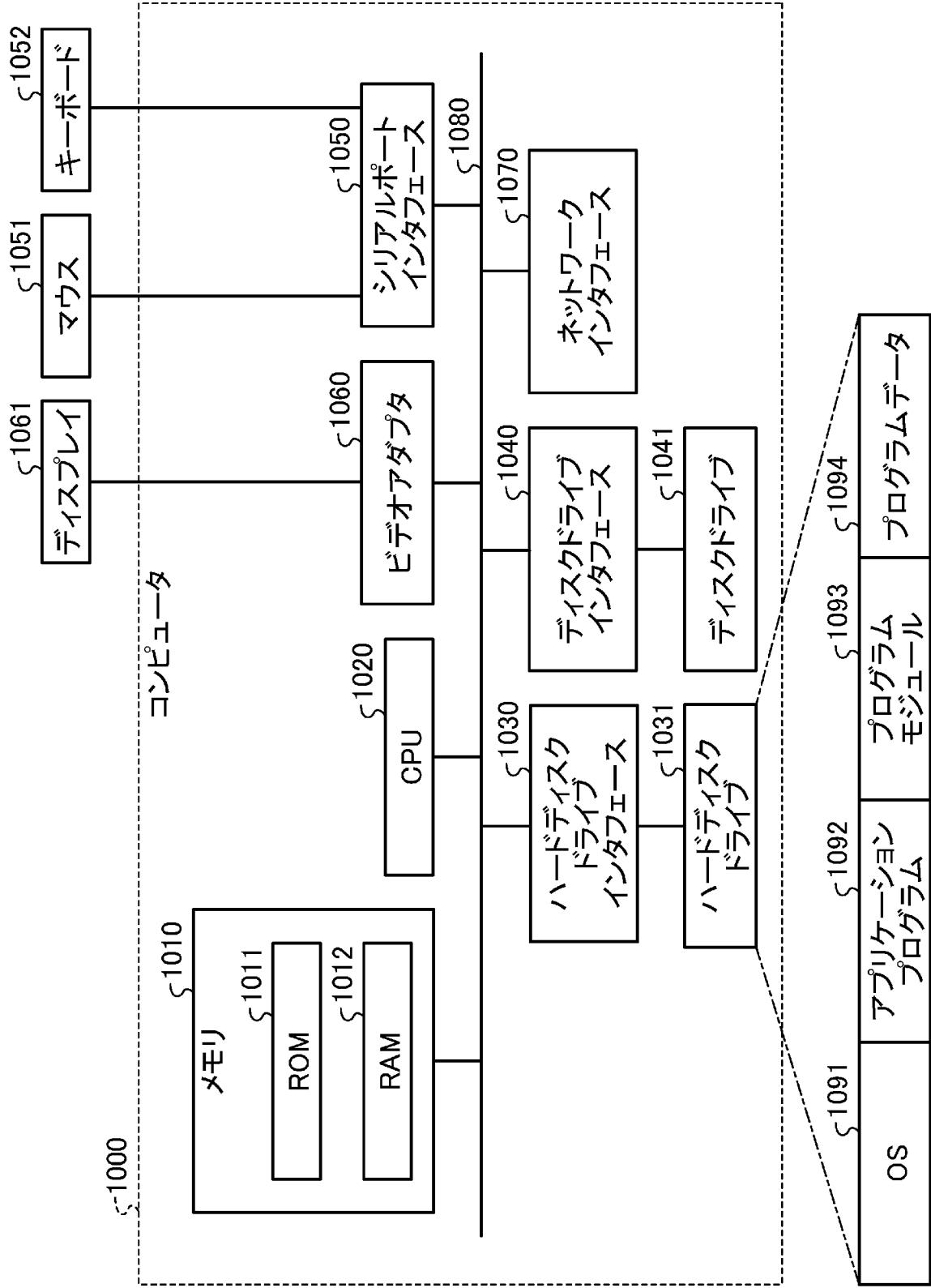
[図13]



[図14]

	Backbone	物体検出 (mAP)	姿勢推定 (mAP)	推論時間 (5000枚)	学習時間
Deformable DETR	R-50	0.423		372秒 1019秒	23時間 49時間
PETR	R-50		0.679	647秒	26時間
本発明のモデル	R-50	0.385	0.667	737秒	35時間
	Swin-T	0.431	0.684	870秒	41時間

[図15]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2023/029898

A. CLASSIFICATION OF SUBJECT MATTER*G06N 3/04*(2023.01)i; *G06N 3/08*(2023.01)i

FI: G06N3/04; G06N3/08

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)

G06N3/04; G06N3/08

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

Published examined utility model applications of Japan 1922-1996
 Published unexamined utility model applications of Japan 1971-2023
 Registered utility model specifications of Japan 1996-2023
 Published registered utility model applications of Japan 1994-2023

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	XU, Yangyang et al. Multi-Task Learning with Multi-Query Transformer for Dense Prediction. arXiv [online]. 07 April 2023, pp. 1-13, [retrieved on 25 September 2023], internet:<URL: https://arxiv.org/pdf/2205.14354v4.pdf > pp. 4-6, fig. 4	1-6
A	RANFTL, Rene et al. Vision Transformers for Dense Prediction. arXiv [online]. 24 March 2021, pp. 1-15, [retrieved on 25 September 2023], internet:<URL: https://arxiv.org/pdf/2103.13413v1.pdf > entire text	1-6
A	CARION, Nicolas et al. End-to-End Object Detection with Transformers. arXiv [online]. 28 May 2020, pp. 1-26, [retrieved on 25 September 2023], internet:<URL: https://arxiv.org/pdf/2005.12872v3.pdf > entire text	1-6

☐ 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

“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

25 September 2023

Date of mailing of the international search report

03 October 2023

Name and mailing address of the ISA/JP

Japan Patent Office (ISA/JP)
 3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915
 Japan

Authorized officer

Telephone No.

A. 発明の属する分野の分類（国際特許分類（IPC）） G06N 3/04(2023.01)i; G06N 3/08(2023.01)i FI: G06N3/04; G06N3/08		
B. 調査を行った分野 調査を行った最小限資料（国際特許分類（IPC）） G06N3/04; G06N3/08 最小限資料以外の資料で調査を行った分野に含まれるもの 日本国実用新案公報 1922-1996年 日本国公開実用新案公報 1971-2023年 日本国実用新案登録公報 1996-2023年 日本国登録実用新案公報 1994-2023年 国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）		
C. 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
X	XU, Yangyang et al., Multi-Task Learning with Multi-Query Transformer for Dense Prediction, arXiv [online], 2023.04.07, pp.1-13, [検索日 2023.09.25]、インターネット:<URL:https://arxiv.org/pdf/2205.14354v4.pdf> 第4～6 ページ、図4	1-6
A	RANFTL, Rene et al., Vision Transformers for Dense Prediction, arXiv [online], 2021.03.24, pp.1-15, [検索日 2023.09.25]、インターネット:<URL:https://arxiv.org/pdf/2103.13413v1.pdf> 全文	1-6
A	CARION, Nicolas et al., End-to-End Object Detection with Transformers, arXiv [online], 2020.05.28, pp.1-26, [検索日 2023.09.25]、インターネット:<URL:https://arxiv.org/pdf/2005.12872v3.pdf> 全文	1-6
☐ C欄の続きにも文献が列挙されている。 ☐ パテントファミリーに関する別紙を参照。		
* 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的な技術水準を示すもの “E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの “L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す） “O” 口頭による開示、使用、展示等に言及する文献 “P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献 “T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの “X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの “Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの “&” 同一パテントファミリー文献		
国際調査を完了した日 25.09.2023		国際調査報告の発送日 03.10.2023
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号		権限のある職員（特許庁審査官） 福西 章人 5B 4687 電話番号 03-3581-1101 内線 3545