I888949

(19)中華民國智慧財產局

(12)發明說明書公告本 **(11)證書號數：TW I888949 B**

(45)公告日： 中華民國 114 (2025) 年 07 月 01 日

(21)申請案號：112138531 (22)申請日： 中華民國 112 (2023) 年 10 月 06 日

(51)Int. Cl.： *G06N3/08 (2023.01)* *G06V10/70 (2022.01)*
G06T1/20 (2006.01)

(71)申請人：國立臺灣科技大學 (中華民國) NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY (TW)
臺北市大安區基隆路 4 段 43 號
利凌企業股份有限公司 (中華民國) MERIT LILIN ENT. CO., LTD. (TW)
新北市五股區五工六路 20 號

(72)發明人：花凱龍 HUA, KAI-LUNG (TW) ；陳永耀 CHEN, YUNG-YAO (TW) ；鍾昕燁 JHONG, SIN-YE (TW) ；林詠翔 LIN, YONG-XIANG (TW) ；羅宥鈞 LO, YOU-JYUN (TW) ；胡志剛 HU, CHIH-KANG (TW) ；黃耀邦 HUANG, YAO-BANG (TW) ；許家雄 HSU, CHIA-HSIUNG (TW)

(74)代理人：高玉駿；楊祺雄

(56)參考文獻：

TW	202215367A	CN	112070111A
CN	114898189A	US	2022/0374647A1

審查人員：黃偉程

申請專利範圍項數：7 項 圖式數：7 共 51 頁

(54)名稱

影像物件辨識模型的訓練方法、影像物件辨識模型及電腦可存取的記錄媒體

(57)摘要

一種影像物件辨識模型的訓練方法，由一電腦裝置執行待訓練的一影像物件辨識模型，該電腦裝置依序將每一組訓練用影像輸入該影像物件辨識模型，由該影像物件辨識模型對每一組訓練用影像包含之在同一時間拍攝且內容重疊的一可見光影像與一熱影像，藉由基於像素對齊模組和基於錨框對齊模組改善雙影像因視野不同造成之特徵偏移，藉由跨模態特徵強化模組針對雙影像特徵強化不同模態的特徵特點，並融合雙影像的全域特徵和區域特徵，而達到多物件偵測效果的提升。

指定代表圖：

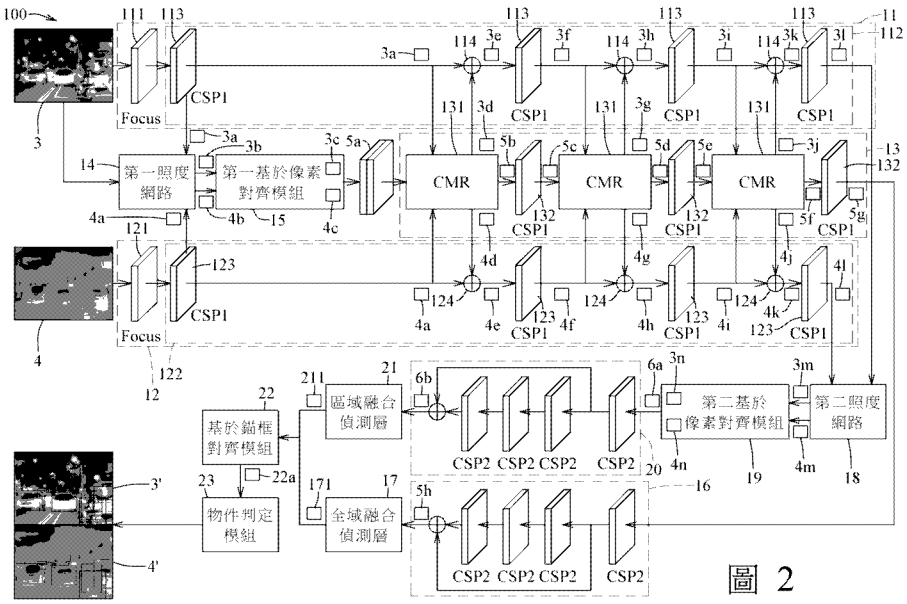

圖 2

符號簡單說明：

100:影像物件辨識模型

11:第一骨幹層

111:聚焦層

112:第一跨階段局部網路(第一 CSPNet)

113:第一跨階段局部模組(第一 CSP 模組)

114:加法器

12:第二骨幹層

121:聚焦層

122:第二跨階段局部網路(第二 CSPNet)

123:第二 CSP 模組

124:加法器

13:第三骨幹層

131:跨模態特徵強化(CMR)模組

132:第三跨階段局部模組(第三 CSP 模組)

14:第一照度網路

15:第一基於像素對齊模組

16:第一多尺度層

17:全域融合偵測層

171:第一候選特徵圖

18:第二照度網路

19:第二基於像素對齊模組

20:第二多尺度層

21:區域融合偵測層

211:第二候選特徵圖

22:基於錨框對齊模組

22a:最終特徵圖

23:物件判定模組

3、3':可見光影像

4、4':熱影像

3a:第一特徵圖

3b:加權後第一特徵圖

3c:校正且加權後第一特徵圖

3d:第一強化特徵圖
3e:第一混合特徵圖
3f:第三特徵圖
3g:第三強化特徵圖
3h:第三混合特徵圖
3i:第五特徵圖
3j:第五強化特徵圖
3k:第五混合特徵圖
3l:第七特徵圖
3m:加權後第三特徵圖
3n:校正且加權後第三特徵圖
4a:第二特徵圖
4b:加權後第二特徵圖
4c:校正且加權後第二特徵圖
4d:第二強化特徵圖
4e:第二混合特徵圖
4f:第四特徵圖
4g:第四強化特徵圖
4h:第四混合特徵圖
4i:第六特徵圖
4j:第六強化特徵圖
4k:第六混合特徵圖
4l:第八特徵圖
4m:加權後第四特徵圖
4n:校正且加權後第四特徵圖
5a:第一融合特徵圖
5b:第一強化融合特徵圖
5c:第三融合特徵圖
5d:第二強化融合特徵圖
5e:第四融合特徵圖
5f:第三強化融合特徵圖
5g:第五融合特徵圖
5h:第一最終融合特徵圖
6a:第五融合特徵圖

6b:第二最終融合特徵圖

I888949

【發明摘要】

【中文發明名稱】 影像物件辨識模型的訓練方法、影像物件辨識模型及電腦可存取的記錄媒體

【中文】

一種影像物件辨識模型的訓練方法，由一電腦裝置執行待訓練的一影像物件辨識模型，該電腦裝置依序將每一組訓練用影像輸入該影像物件辨識模型，由該影像物件辨識模型對每一組訓練用影像包含之在同一時間拍攝且內容重疊的一可見光影像與一熱影像，藉由基於像素對齊模組和基於錨框對齊模組改善雙影像因視野不同造成之特徵偏移，藉由跨模態特徵強化模組針對雙影像特徵強化不同模態的特徵特點，並融合雙影像的全域特徵和區域特徵，而達到多物件偵測效果的提升。

【指定代表圖】：圖 2

【代表圖之符號簡單說明】

100：影像物件辨識模型
11：第一骨幹層
111：聚焦層
112：第一跨階段局部網路(第一 CSPNet)
113：第一跨階段局部模組(第一 CSP 模組)
114：加法器
12：第二骨幹層
121：聚焦層
122：第二跨階段局部網路(第二 CSPNet)
123：第二 CSP 模組
124：加法器
13：第三骨幹層
131：跨模態特徵強化(CMR)模組
132：第三跨階段局部模組(第三 CSP 模組)
14：第一照度網路
15：第一基於像素對齊模組
16：第一多尺度層
17：全域融合偵測層

171：第一候選特徵圖

18：第二照度網路

19：第二基於像素對齊模組

20：第二多尺度層

21：區域融合偵測層

211：第二候選特徵圖

22：基於錨框對齊模組

22a：最終特徵圖

23：物件判定模組

3、3’：可見光影像

4、4’：熱影像

3a：第一特徵圖

3b：加權後第一特徵圖

3c：校正且加權後第一特徵圖

3d：第一強化特徵圖

3e：第一混合特徵圖

3f：第三特徵圖

3g：第三強化特徵圖

3h：第三混合特徵圖

3i：第五特徵圖

3j：第五強化特徵圖

3k：第五混合特徵圖

3l：第七特徵圖

3m：加權後第三特徵圖

3n：校正且加權後第三特徵圖

4a：第二特徵圖

4b：加權後第二特徵圖

4c：校正且加權後第二特徵圖

4d：第二強化特徵圖

4e：第二混合特徵圖

4f：第四特徵圖

4g：第四強化特徵圖

4h：第四混合特徵圖

4i：第六特徵圖

4j：第六強化特徵圖

4k：第六混合特徵圖

4l：第八特徵圖

4m：加權後第四特徵圖

4n：校正且加權後第四特徵圖

5a：第一融合特徵圖

5b：第一強化融合特徵圖

5c：第三融合特徵圖

5d：第二強化融合特徵圖

5e：第四融合特徵圖

5f：第三強化融合特徵圖

5g：第五融合特徵圖

5h：第一最終融合特徵圖

6a：第五融合特徵圖

6b：第二最終融合特徵圖

【發明說明書】

【中文發明名稱】 影像物件辨識模型的訓練方法、影像物件辨識模型及電腦可存取的記錄媒體

【技術領域】

【0001】 本發明是有關於一種影像物件辨識模型及其訓練方法，特別是指一種能根據同一成像時間獲得之同一場景的熱影像與可見光影像進行影像物件辨識的影像物件辨識模型及其訓練方法。

【先前技術】

【0002】 可見光相機(RGB Camera)在天候良好、光線明亮時，其拍攝範圍內之物件成像效果良好，但在光線昏暗，如夜晚無光源處，其成像效果則與光線強弱成反比。而在雨、雪、霧等天候不良或有煙、塵的環境時，則易遭遮蔽且無法穿透，成像效果不佳，以致影響辨識影像中之物件的識別率。熱感攝影機(或稱紅外線相機，Thermal Camera)在天候不佳或光線昏暗環境下，其成像效果較可見光相機佳，但熱感攝影機僅能描繪物件的外型，不能顯示物件的細節輪廓，例如無法顯示人臉的細部特徵，且當所拍攝的相鄰物件溫度相近時，熱感攝影機易混淆相鄰物件而影響辨識影像中之物件的識別率。

【0003】 因此，為解決上述問題，傳統採用上述兩種影像進行影

像中之物件辨識的方法會設定一個切換機制，例如白天使用可見光相機拍攝的可見光影像進行物件辨識，晚上則切換至使用熱感攝影機拍攝的熱影像進行物件辨識；但此種做法需要特別考慮時段而且過度依賴單一種影像，例如即使在晚上但燈火通明的地方，可見光影像的成像效果未必較熱影像差，反之，即使在晚上但溫度差異不大的環境，例如冬天或冰天雪地的地方，熱影像的成像效果亦不見得較可見光影像佳。

【0004】 因此，若能同時採用上述兩種影像進行影像物件辨識，可利用影像互補的效果，而不需考量時段或環境的變化對應切換不同的影像辨識機制，並可進行全天候的影像辨識。

【發明內容】

【0005】 因此，本發明之目的，即在提供一種影像物件辨識模型的訓練方法、被該方法訓練的一種影像物件辨識模型以及儲存該影像物件辨識模型的一種電腦可存取的記錄媒體，該影像物件辨識模型同時採用內容重疊的熱影像與可見光影像進行影像物件辨識，利用影像互補的效果，達到全天候影像辨識。

【0006】 於是，本發明一種影像物件辨識模型的訓練方法，包括：A、一電腦裝置的一處理單元執行預先載入之待訓練的一影像物件辨識模型，該影像物件辨識模型包括一第一骨幹層、一第二骨幹層、一第三骨幹層、與該第三骨幹層串接的一跨模態特徵強化模組、與

該第一骨幹層和該第二骨幹層連接的一第一照度網路、與該第一照度網路和該跨模態特徵強化模組連接的一第一基於像素對齊模組、與該第三骨幹層連接的一第一多尺度層、與該第一多尺度層連接的一全域融合偵測層、與該第一骨幹層和該第二骨幹層連接的一第二照度網路、與該第二照度網路連接的一第二基於像素對齊模組、與該第二基於像素對齊模組連接的一第二多尺度層、與該第二多尺度層連接的一區域融合偵測層、與該全域融合偵測層和該區域融合偵測層連接的一基於錨框對齊模組以及與該基於錨框對齊模組連接的一物件判定模組；及 B、該電腦裝置提供複數組訓練用影像給該處理單元，每一組訓練用影像包含在同一時間拍攝且內容重疊的一可見光影像與一熱影像；該處理單元依序將每一組訓練用影像輸入該影像物件辨識模型，以藉由反覆執行下述動作訓練該影像物件辨識模型。

【0007】 B1、將該可見光影像和該熱影像各別對應輸入該第一骨幹層和該第二骨幹層，並將該可見光影像輸入該第一照度網路和該第二照度網路，使該第一骨幹層對該可見光影像進行區域特徵提取並輸出一第一特徵圖至該第一照度網路和該跨模態特徵強化模組，並使該第二骨幹層對該熱影像進行區域特徵提取並輸出一第二特徵圖至該第一照度網路和該跨模態特徵強化模組；且該第一照度網路和該第二照度網路根據該可見光影像求得與明亮環境相關的

一第一權重和與陰暗環境相關的一第二權重。

【0008】 B2、該第一照度網路將該第一特徵圖以該第一權重加權以產生一加權後第一特徵圖，且將該第二特徵圖以該第二權重加權以產生一加權後第二特徵圖，並將該加權後第一特徵圖和該加權後第二特徵圖輸出至該第一基於像素對齊模組。

【0009】 B3、該第一基於像素對齊模組根據該加權後第一特徵圖的特徵像素和該加權後第二特徵圖的特徵像素各自與一偏移場域之間的一偏移量，校正該加權後第一特徵圖和該加權後第二特徵圖，以產生特徵像素對齊的一校正且加權後第一特徵圖和一校正且加權後第二特徵圖，並將該校正且加權後第一特徵圖與該校正且加權後第二特徵圖相疊合而產生並輸出一第一融合特徵圖至該跨模態特徵強化模組。

【0010】 B4、該跨模態特徵強化模組利用基於移位窗口的自注意力機制對輸入的該第一特徵圖、該第二特徵圖和該第一融合特徵圖進行全域特徵擷取，以對應產生三個強化特徵圖，並分別輸出三個強化特徵圖至相對應的該第一骨幹層、該第二骨幹層和該第三骨幹層。

【0011】 B5、該第三骨幹層對該跨模態特徵強化模組輸出的該強化特徵圖進行特徵擷取而產生並輸出一特徵圖至該第一多尺度層；該第一骨幹層基於該第一特徵圖與該跨模態特徵強化模組輸出

的該強化特徵圖進行特徵提取以產生並輸出一特徵圖至該第二照度網路，且該第二骨幹層基於該第二特徵圖與該跨模態特徵強化模組輸出的該強化特徵圖進行特徵提取以產生並輸出一特徵圖至該第二照度網路。

【0012】 B6、該第一多尺度層對該第三骨幹層輸出的該特徵圖進行基於不同尺度的特徵擷取，以產生並輸出一特徵圖至該全域融合偵測層；該全域融合偵測層根據該第一多尺度層輸出的該特徵圖中的影像特徵進行候選框偵測和物件辨識，以產生並輸出具有複數個第一候選物件資訊的一特徵圖至該基於錨框對齊模組。

【0013】 B7、該第二照度網路將該第一骨幹層輸出的該特徵圖以該第一權重加權以產生一加權後特徵圖，且將該第二骨幹層輸出的該特徵圖以該第二權重加權以產生一加權後特徵圖，並將該二個加權後特徵圖輸出至該第二基於像素對齊模組。

【0014】 B8、該第二基於像素對齊模組根據該第二照度網路輸出的該二個加權後特徵圖的特徵像素各自與一偏移場域之間的一偏移量，校正該二個加權後特徵圖，以產生特徵像素對齊的二個校正且加權後特徵圖，並將該二個校正且加權後特徵圖相疊合而產生並輸出一融合特徵圖至該第二多尺度層；該第二多尺度層對該第二基於像素對齊模組輸出的該融合特徵圖進行基於不同尺度的特徵擷取，以產生並輸出一特徵圖至該區域融合偵測層；該區域融合偵

測層根據該第二多尺度層輸出的該特徵圖中的影像特徵進行候選框偵測和物件辨識，以產生並輸出具有複數個第二候選物件框的一特徵圖至該基於錨框對齊模組。

【0015】 B9、該基於錨框對齊模組根據輸入的二個特徵圖對應的該等第一候選物件框和該等第二候選物件框的相對偏移位置，校正該等第一候選物件框和該等第二候選物件框，使校正後的該等第一候選物件框和校正後的該等第二候選物件框及其所涵蓋的影像物件對齊，以此產生並輸出具有複數個最終候選物件框的一特徵圖至該物件判定模組。

【0016】 B10、該物件判定模組根據該基於錨框對齊模組輸出的該特徵圖中的該等最終候選物件框所對應的信心指數，從該等最終候選物件框選出最佳的候選物件框，並將選出的最佳候選物件框顯示在該可見光影像中。

【0017】 在本發明的一些實施態樣中，每一組訓練用影像包含的該熱影像是預先以一座標投影矩陣校正的校正後熱影像，使該校正後熱影像能與相對應的該可見光影像對齊，且該座標投影矩陣是根據該熱影像與相對應的該可見光影像兩者的視野差異而求得。

【0018】 在本發明的一些實施態樣中，該跨模態特徵強化模組包含依序串連的四個特徵強化層，該等特徵強化層包含用於正規化輸入的特徵的一正規化層；第一個特徵強化層包含的 W-MSA 層和

第三個特徵強化層包含的SW-MSA層對輸入的特徵進行多頭自注意力處理以提取特徵；第二個和第四個特徵強化層包含的一卷積層對輸入的特徵進行卷積運算以提取特徵。

【0019】 在本發明的一些實施態樣中，該基於錨框對齊模組包含依序串連的一特徵融合層、一錨框位移層及一卷積層，該特徵融合層將輸入的二個特徵圖相疊合成為一融合特徵圖，再由該錨框位移層根據該融合特徵圖中該等第一候選物件框和該等第二候選物件框的相對偏移位置，校正該等第一候選物件框和該等第二候選物件框，使校正後的該等第一候選物件框和校正後的該等第二候選物件框及其所涵蓋的影像物件對齊，再由該卷積層對該融合特徵圖進行卷積運算以降維，並輸出具有該等最終候選物件框的該特徵圖。

【0020】 在本發明的一些實施態樣中，該跨模態特徵強化模組是第一個跨模態特徵強化模組，且該影像物件辨識模型還包括與第一個跨模態特徵強化模組和該第三骨幹層串接的第二個跨模態特徵強化模組和第三個跨模態特徵強化模組；且步驟B5還包括下列步驟。

【0021】 B51、該第一骨幹層基於該第一特徵圖與該強化特徵圖進行特徵提取而產生的該特徵圖被輸入第二個跨模態特徵強化模組，且該第二骨幹層基於該第二特徵圖與該強化特徵圖進行特徵提取而產生的該特徵圖被輸入第二個跨模態特徵強化模組，該第三骨

幹層對該強化特徵圖進行特徵擷取而產生的該特徵圖被輸入第二個跨模態特徵強化模組。

【0022】 B52、第二個跨模態特徵強化模組對輸入的三個特徵圖進行特徵強化，以對應產生三個強化特徵圖，並分別輸出該三個強化特徵圖至相對應的該第一骨幹層、該第二骨幹層和該第三骨幹層。

【0023】 B53、該第一骨幹層基於輸入該第二個跨模態特徵強化模組的該特徵圖和該第二個跨模態特徵強化模組輸出的強化特徵圖進行特徵提取而產生並輸出一特徵圖至該第三個跨模態特徵強化模組；該第二骨幹層基於輸入該第二個跨模態特徵強化模組的該特徵圖和該第二個跨模態特徵強化模組輸出的強化特徵圖進行特徵提取而產生並輸出一特徵圖至該第三個跨模態特徵強化模組；該第三骨幹層對第二個跨模態特徵強化模組輸出的強化特徵圖進行特徵提取而產生並輸出一特徵圖至該第三個跨模態特徵強化模組。

【0024】 B54、第三個跨模態特徵強化模組對輸入的三個特徵圖進行特徵強化，以對應產生三個強化特徵圖，並分別輸出三個強化特徵圖至相對應的該第一骨幹層、該第二骨幹層和該第三骨幹層。

【0025】 B55、該第三骨幹層對第三個跨模態特徵強化模組輸出的強化特徵圖進行特徵提取而產生並輸出該特徵圖至該第一多尺度層；該第一骨幹層基於輸入該第三個跨模態特徵強化模組的該特徵圖和該第三個跨模態特徵強化模組輸出的強化特徵圖進行特徵

提取而產生並輸出該特徵圖至該第二照度網路；該第二骨幹層基於輸入該第三個跨模態特徵強化模組的該特徵圖和該第三個跨模態特徵強化模組輸出的強化特徵圖進行特徵提取而產生並輸出該特徵圖至該第二照度網路。

【0026】 此外，本發明一種影像物件辨識模型，其係根據上述影像物件辨識模型的訓練方法訓練而成，而能接受包含在同一時間拍攝且內容重疊的一待辨識可見光影像與一待辨識熱影像的一組待辨識影像，以根據該待辨識可見光影像與該待辨識熱影像辨識出該待辨識可見光影像中的物件。

【0027】 再者，本發明一種電腦可存取的記錄媒體，其中存有如上所述影像物件辨識模型的訓練方法中記載的該影像物件辨識模型，且該影像物件辨識模型藉由如上所述影像物件辨識模型的訓練方法訓練完成後，能根據輸入之在同一時間拍攝且內容重疊的一可見光影像與一熱影像辨識該可見光影像中的物件。

【0028】 本發明之功效在於：藉由照度網路根據可見光影像呈現的照度調整可見光影像的特徵圖和熱影像的特徵圖的比重，並藉由基於像素對齊模組和基於錨框對齊模組改善雙影像因視野不同造成之特徵偏移；且藉由跨模態特徵強化模組針對雙影像特徵的不平衡，套用自注意力機制強化不同模態的特徵特點，改善模型經過訓練後的特徵丟失，並融合全域特徵和區域特徵，進行全天候的影

像辨識並提升物件辨識率。

【圖式簡單說明】

【0029】 本發明之其他的特徵及功效，將於參照圖式的實施方式中清楚地顯示，其中：

圖 1 是本發明影像物件辨識模型的訓練方法的第一實施例的主要流程；

圖 2 是第一實施例的影像物件辨識模型的架構方塊示意圖；

圖 3A~圖 3D 是第一實施例訓練影像物件辨識模型的細部流程；及

圖 4 是第一實施例的 CMR 模組的細部架構方塊示意圖；

圖 5 是第一實施例的基於錨框對齊模組的細部架構方塊示意圖；

圖 6 是本發明的影像物件辨識模型的第二實施例的架構方塊示意圖；及

圖 7 是本發明的影像物件辨識模型的第三實施例的架構方塊示意圖。

【實施方式】

【0030】 在本發明被詳細描述之前，應當注意在以下的說明內容中，類似的元件是以相同的編號來表示。

【0031】 參閱圖 1 所示，是本發明影像物件辨識模型的訓練方

法的一第一實施例的主要流程步驟，首先，如圖 1 的步驟 S1，本實施例由一電腦裝置(圖未示)的一處理單元(例如中央處理器)執行預先載入之待訓練的一影像物件辨識模型 100，該影像物件辨識模型 100 是儲存在一電腦電存取的記錄媒體，例如該電腦裝置的記憶體模組或硬碟中的軟體程式，且該影像物件辨識模型 100 可以是基於 YOLO 系列，例如 YOLOv2、YOLOv3、YOLOv4、YOLOv5、YOLOv7、YOLOv8、YOLOR、ScaledYOLOv4 等至少其中之一進行開發的物件偵測模型，本實施例的該影像物件辨識模型 100 將以基於 YOLOv5 開發 1 的架構進行說明。

【0032】 如圖 2 所示，該影像物件辨識模型 100 主要包括第一骨幹(backbone)層 11、一第二骨幹層 12、一第三骨幹層 13、與該第三骨幹層 13 串接的三個跨模態特徵強化(Cross-modality Reinforcement Module，以下簡稱 CMR)模組 131、一第一照度網路(Illumination Mechanism)14、一第一基於像素對齊(Pixel-based Alignment)模組 15、一第一多尺度層 16、一全域融合偵測層(Global Fusion Detection Head)17、一第二照度網路 18、一第二基於像素對齊模組 19、一第二多尺度層 20、一區域融合偵測層(Local Fusion Detection Head)21、一基於錨框對齊(Anchor-based Alignment)模組 22 及一物件判定模組 23。

【0033】 該第一骨幹層 11 包含一聚焦(Focus)層 111 和一第一

跨階段局部網路 (Cross Stage Partial Network，以下簡稱CSPNet)112；該聚焦(Focus)層 111 主要對輸入的圖像進行切片及堆疊，類似於鄰近下採樣(縮小圖像)，再對得到的新圖片經過卷積操作，以得到沒有信息丟失的下採樣特徵圖，並輸出特徵圖至該第一 CSPNet112。CSPNet 是 YOLO 的現有技術且非本案重點所在，故在此不予詳述。

【0034】 該第一 CSPNet112 的主要目的是使網路架構能夠獲取更豐富的梯度融合信息並降低計算量，具體而言，該第一 CSPNet112 包含複數個串連的第一跨階段局部模組(CSP1)113(以下簡稱第一 CSP 模組 113)和連接在相鄰的兩兩第一跨階段局部(CSP1)模組 113 之間的複數個加法器 114。本實施例是以該第一 CSPNet112 包含四個第一 CSP 模組 113 和三個連接相鄰的兩兩第一 CSP 模組 113 的加法器 114 為例。

【0035】 該第二骨幹層 12 具有和該第一骨幹層 12 相同的架構，而同樣具有一聚焦(Focus)層 121 和一第二跨階段局部網路(Cross Stage Partial Network，以下簡稱 CSPNet)122，並以該第二 CSPNet122 包含四個第二 CSP 模組 123 和三個連接相鄰的兩兩第二 CSP 模組 123 的加法器 124 為例。

【0036】 值得一提的是，上述的該聚焦(Focus)層 111、121 並非必要，也可以視實際應用情況被省略。

【0037】 在本實施例中，該第三骨幹層 13 包含與各 CMR 模組 131 的輸出端連接的三個第三跨階段局部(CSP)模組 132。其中，第一個 CMR 模組 131 與該第一骨幹層 11 中的第一個該第一 CSP 模組 113 的輸出端和第一個該第一加法器 114 連接，並與該第二骨幹層 12 中的第一個該第二 CSP 模組 123 的輸出端和第一個該第二加法器 124 連接；第二個 CMR 模組 131 與該第一骨幹層 11 中的第二個該第一 CSP 模組 113 的輸出端和第二個該第一加法器 114 連接，並與該第二骨幹層 12 中的第二個該第二 CSP 模組 123 的輸出端和第二個該第二加法器 124 連接；第三個 CMR 模組 131 與該第一骨幹層 11 中的第三個該第一 CSP 模組 113 的輸出端和第三個該第一加法器 114 連接，並與該第二骨幹層 12 中的第三個該第二 CSP 模組 123 的輸出端和第三個該第二加法器 124 連接。

【0038】 該第一照度網路 14 與該第一骨幹層 11 中的第一個該第一 CSP 模組 113 的輸出端和該第二骨幹層 12 中的第一個該第二 CSP 模組 123 的輸出端連接。

【0039】 該第一基於像素對齊模組 15 與該第一照度網路 14 的輸出端和第一個 CMR 模組 131 的輸入端連接。

【0040】 該第一多尺度層 16，又稱頸部(Neck)層，其與該第三骨幹層 13 的最後一個(即第三個)該第三 CSP 模組 132 的輸出端連接。該第一多尺度層 16 也是由多個跨階段局部(CSP2)模組 161

組成。

【0041】 該全域融合偵測層 17 與該第一多尺度層 16 的輸出端連接。

【0042】 該第二照度網路 18 與該第一骨幹層 11 的輸出端，即最後一個該第一 CSP 模組 113 的輸出端和該第二骨幹層 12 的輸出端，即最後一個該第二 CSP 模組 123 的輸出端連接。此外，該第一照度網路 14 和該第二照度網路 18 是採用相同的深度學習網路，例如基於卷積神經網路(CNN)的 R-CNN，且皆被預先訓練完成而能夠偵測輸入的一影像中的照度(亮度)以輸出與明亮環境相關(表徵明亮環境，例如白天、晴天)的一第一權重和與陰暗環境相關(表徵陰暗環境，例如夜晚、陰天或隧道)的一第二權重，亦即，該第一權重代表明亮的機率，該第二權重代表陰暗的機率，因此該第一權重和該第二權重的總和為 1。由於照度網路是現有技術且非本案重點所在，故在此不予詳述，有關照度網路(Illumination Mechanism)的技術細節可以參見論文「Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems」其中的「3.2 Illumination Aware Feature Alignment Module」以及論文「Illumination-aware Faster R-CNN for Robust Multispectral Pedestrian Detection」。

【0043】 該第二基於像素對齊模組 19 與該第二照度網路 18 的

輸出端連接。且該第二基於像素對齊模組 19 與該第一基於像素對齊模組 15 相同。

【0044】 該第二多尺度層 20 與該第二基於像素對齊模組 19 的輸出端連接。且該第二多尺度層 20 具有與該第一多尺度層 16 相同的組成架構。

【0045】 該區域融合偵測層 21 與該第二多尺度層 20 的輸出端連接。該區域融合偵測層 21 與該全域融合偵測層 17 具有相同的架構，且都是基於 YOLOv5 的 Head 架構開發。

【0046】 該基於錨框對齊模組 22 與該全域融合偵測層 17 的輸出端和該區域融合偵測層 21 的輸出端連接。

【0047】 該物件判定模組 23 與該基於錨框對齊模組 22 的輸出端連接。該物件判定模組 23 在本實施例中是採用 DIOU-NMS 演算法，其中 DIOU 的全文為 Distance Intersection over Union，NMS 的全文為 Non-Max Suppression，而 DIOU-NMS 演算法的主要原理為利用信心指數來判斷輸入的多個物件候選框其中哪些是最佳的候選框。且由於 DIOU-NMS 演算法已是一習知演算法，且非本案主要重點所在，故在此不予詳述。

【0048】 此外，上述的 CSPNet、多尺度層和該物件判定模組 23 的具體細部架構並非本案技術重點所在，可參見上述公開的 YOLO 系列的相關文獻或介紹，故在此不予贅述。

【0049】 然後，如圖 1 的步驟 S2，該電腦裝置提供複數組訓練用影像給該處理單元，每一組訓練用影像包含在同一時間拍攝且內容重疊的一可見光影像與一熱影像；具體而言，該等訓練用影像例如是由設置在一車輛上的一影像擷取系統收集，該影像擷取系統包含並排地固定在車輛前面的一可見光相機及一熱影像相機，以透過該可見光相機和該熱影像同時拍攝可見光影像和熱影像；此外，該影像擷取系統還事先基於該可見光相機的視野(或視角，簡稱 FOV)與該熱影像相機的視野兩者之間的差異，計算出用以校正熱影像的一座標投影矩陣，並以該座標投影矩陣校正該熱影像相機拍攝的熱影像，使校正後熱影像的涵蓋範圍能與相對應的可見光影像重疊，而能與相對應的可見光影像對齊，因此，每一組訓練用影像中的熱影像都是經過校正的熱影像。

【0050】 接著，如圖 1 的步驟 S3，該處理單元依序將每一組訓練用影像輸入該影像物件辨識模型 100，以藉由圖 3A~圖 3D 所示的流程和下述訓練過程訓練該影像物件辨識模型 100。

【0051】 首先，如圖 3A 的步驟 S31，該可見光影像 3 和該熱影像 4 各別對應輸入該第一骨幹層 11 和該第二骨幹層 12，使該第一骨幹層 11 的第一個該第一 CSP 模組 113 對該聚焦(Focus)層 111 輸出的一初始特徵圖進行區域(Local)特徵擷取，以產生並分別輸出一第一特徵圖 3a 至該第一照度網路 14、第一個 CMR 模

組 131 及第一個該第一加法器 114，並使該第二骨幹層 12 的第一個該第二 CSP 模組 123 對該聚焦(Focus)層 121 輸出的一初始特徵圖進行區域特徵擷取，以產生並分別輸出一第二特徵圖 4a 至該第一照度網路 14、第一個 CMR 模組 131 及第一個該第二加法器 124。

【0052】 同時，該第一照度網路 14 和該第二照度網路 18 根據輸入的(每一組訓練用影像中的)該可見光影像求得與明亮環境相關的該第一權重和與陰暗環境相關的該第二權重。

【0053】 然後，如圖 3A 的步驟 S32，該第一照度網路 14 將輸入的該第一特徵圖 3a 以該第一權重加權以產生一加權後第一特徵圖 3b，且將該第二特徵圖 4a 以該第二權重加權以產生一加權後第二特徵圖 4b，並將該加權後第一特徵圖 3b 和該加權後第二特徵圖 4b 輸出至該第一基於像素對齊模組 15。

【0054】 接著，如圖 3 的步驟 S33，該第一基於像素對齊(校正)模組 15 根據輸入的該加權後第一特徵圖 3b 的特徵像素和該加權後第二特徵圖 4b 的特徵像素各自與一偏移場域(offset field)之間的一偏移量(offset)，校正該加權後第一特徵圖 3b 的特徵像素和該加權後第二特徵圖 4b 的特徵像素的位置，而產生特徵像素對齊的一校正且加權後第一特徵圖 3c 和一校正且加權後第二特徵圖 4c，並將該校正且加權後第一特徵圖 3c 與該校正且加權後第二特徵圖

4c 相疊合而產生並輸出一第一融合特徵圖 5a 至該第三骨幹層 13 的第一個 CMR 模組 131。

【0055】 其中，該偏移場域(offset field) 是透過額外的卷積層學習該加權後第一特徵圖 3b 的特徵像素要如何偏移以及該加權後第二特徵圖 4b 的特徵像素要如何偏移，才會讓該加權後第一特徵圖 3b 和該加權後第二特徵圖 4b 的特徵像素對齊，並藉此得到與該加權後第一特徵圖 3b 的特徵像素和該加權後第二特徵圖 4b 的特徵像素對應的複數個偏移(校正)參考點，然後，該第一基於像素對齊(校正)模組 15 根據該偏移場域中的該等偏移參考點與該加權後第一特徵圖 3b 中相對應的每一特徵像素之間的一偏移量(offset)以及該等偏移參考點與該加權後第二特徵圖 4b 的每一特徵像素之間的一偏移量，校正該加權後第一特徵圖 3b 的特徵像素和該加權後第二特徵圖 4b 的特徵像素的位置，其技術細節可以參見「Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems」該篇論文第 8 頁對於 Fig.4 的說明和第 9 頁第 1 段說明，以及 https://zhuanlan.zhihu.com/p/52476083 的記載內容。

【0056】 接著，如圖 3A 的步驟 S34，第一個 CMR 模組 131 利用基於移位窗口的自注意力(self-attention)機制對輸入的該第一特徵圖 3a、該第二特徵圖 4a 和該第一融合特徵圖 5a 進行全域特

徵擷取，以強化不同模態的特徵特點，改善模型經過訓練後的特徵丟失，而對應產生一第一強化特徵圖 3d、一第二強化特徵圖 4d 及一第一強化融合特徵圖 5b，並輸出該第一強化特徵圖 3d 至該第一骨幹層 11 的第一個該第一加法器 114，輸出該第二強化特徵圖 4d 至該第二骨幹層 12 的第一個該第二加法器 124，輸出該第一強化融合特徵圖 5b 至該第三骨幹層 13 的第一個該第三 CSP 模組 132。

【0057】 具體而言，如圖 4 所示，CMR 模組 131 是參考 swim Transformer 架構開發，並包括透過加法器 133 依序串連的一第一特徵強化層 134、一第二特徵強化層 135、一第三特徵強化層 136 和一第四特徵強化層 137；且 CMR 模組 131 會先透過 Patch Embedding 處理，將二維的該第一特徵圖 3a、該第二特徵圖 4a 和該第一融合特徵圖 5a 轉換成一維的一第一特徵陣列 41，再將該第一特徵陣列 41 輸入該第一特徵強化層 134。其中，Patch Embedding 是一種將二維影像切分為一維圖塊向量的演算法，其技術細節可以參見「AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE」此一論文。

【0058】 其中第一至第四特徵強化層 134~137 中的 LN(Layer Normalization，層正規化)層主要用於正規化特徵，以確保輸出的特徵數值在同一分布；第一特徵強化層 134 中的 W-MSA 層和

第三特徵強化層 136 中的 SW-MSA 層如同原始論文記載，主要都是對輸入的特徵進行多頭自注意力(Multi-headed Self-Attention)處理，其處理目的皆為特徵提取。其中 W-MSA 層是對 windows 內的特徵做自注意力(Self-Attention)處理，SW-MSA 層是先對 windows 做平移，再將新的 windows 內的特徵做自注意力(Self-Attention)處理，有關 windows 如何進行及平移(shift)可以參考 swim Transformer 的原始論文「Swin Transformer：Hierarchical Vision Transformer using Shifted Windows」。而第二和第四特徵強化層 135、137 中的 conv(Convolutional，卷積)層則是對輸入的特徵進行卷積運算以提取特徵，且本實施例以 conv 層取代 swim Transformer 中的 MLP 層的主要原因是經由實驗得知 conv 層會讓物件偵測(辨識)結果更好。

【0059】 且該第一特徵陣列 41 經過第一至第四特徵強化層 134~137 處理後，該第四特徵強化層 137 會產生並輸出一維的一第二特徵陣列 42，CMR 模組 131 再將該第二特徵陣列 42 轉換成二維的該第一強化特徵圖 3d、該第二強化特徵圖 4d 及該第一強化融合特徵圖 5b 並輸出。

【0060】 接著，如圖 3B 的步驟 S35，第一個該第一加法器 114 將第一個 CMR 模組 131 輸出的該第一強化特徵圖 3d 與該第一特

徵圖 3a 相加，以產生並輸出一第一混合特徵圖 3e 至第二個該第一 CSP 模組 113；且第一個該第二加法器 124 將第一個 CMR 模組 131 輸出的該第二強化特徵圖 4d 與該第二特徵圖 4a 相加，以產生並輸出一第二混合特徵圖 4e 至第二個該第二 CSP 模組 123。第二個該第一 CSP 模組 113 對該第一混合特徵圖 3e 進行特徵提取以產生並輸出一第三特徵圖 3f 至第二個該第一加法器 114 和第二個 CMR 模組 131，且第二個該第二 CSP 模組 123 對該第二混合特徵圖 4e 進行特徵提取以產生並輸出一第四特徵圖 4f 至第二個該第二加法器 114 和第二個 CMR 模組 131，且該第三骨幹層 13 的第一個該第三 CSP 模組 132 對該第一強化融合特徵圖 5b 再次進行特徵擷取而產生並輸出一第三融合特徵圖 5c 至第二個 CMR 模組 131。

【0061】 接著，如圖 3B 的步驟 S36，第二個 CMR 模組 131 重覆第一個 CMR 模組 131 的動作，利用基於移位窗口的自注意力(self-attention)機制對第二個該第一 CSP 模組 113 輸出的一第三特徵圖 3f、第二個該第二 CSP 模組 123 輸出的一第四特徵圖 4f 以及第一個該第三 CSP 模組 132 輸出的第三融合特徵圖 5c 進行特徵強化，以對應產生一第三強化特徵圖 3g、一第四強化特徵圖 4g 及一第二強化融合特徵圖 5d，並輸出該第三強化特徵圖 3g 至該第一骨幹層 11 的第二個該第一加法器 114，輸出該第四強化特徵圖 4g 至該第二骨幹層 12 的第二個該第二加法器 124，輸出該第

二強化融合特徵圖 5d 至該第三骨幹層 13 的第二個該第三 CSP 模組 132。

【0062】 接著，如圖 3B 的步驟 S37，第二個該第一加法器 114 將第二個 CMR 模組 131 輸出的該第三強化特徵圖 3g 與該第三特徵圖 3f 相加，以產生並輸出一第三混合特徵圖 3h 至第三個該第一 CSP 模組 113；且第二個該第二加法器 124 將第二個 CMR 模組 131 輸出的該第四強化特徵圖 4g 與該第四特徵圖 4f 相加，以產生並輸出一第四混合特徵圖 4h 至第三個該第二 CSP 模組 123。且第三個該第一 CSP 模組 113 對該第三混合特徵圖 3h 進行特徵提取以產生並輸出一第五特徵圖 3i 至第三個該第一加法器 114 和第三個 CMR 模組 131；且第三個該第二 CSP 模組 123 對該第四混合特徵圖 4h 進行特徵提取以產生並輸出一第六特徵圖 4i 至第三個該第二加法器 114 和第三個 CMR 模組 131；且該第三骨幹層 13 的第二個該第三 CSP 模組 132 對該第二強化融合特徵圖 5d 再次進行特徵擷取而產生並輸出一第四融合特徵圖 5e 至第三個 CMR 模組 131。

【0063】 接著，如圖 3B 的步驟 S38，第三個 CMR 模組 131 同樣地重覆第一個 CMR 模組 131 的動作，利用基於移位窗口的自注意力(self-attention)機制對第三個該第一 CSP 模組 113 輸出的一第五特徵圖 3i、第三個該第二 CSP 模組 123 輸出的一第六特徵

圖 4i 以及第二個該第三 CSP 模組 131 輸出的第四融合特徵圖 5e 進行特徵強化，以對應產生一第五強化特徵圖 3j、一第六強化特徵圖 4j 及一第三強化融合特徵圖 5f，並輸出該第五強化特徵圖 3j 至該第一骨幹層 11 的第三個該第一加法器 114，輸出該第六強化特徵圖 4j 至該第二骨幹層 12 的第三個該第二加法器 124，輸出該第三強化融合特徵圖 5f 至該第三骨幹層 13 的第三個該 CSP 模組 132。

【0064】 接著，如圖 3C 的步驟 S39，第三個該第一加法器 114 將第三個 CMR 模組 131 輸出的該第五強化特徵圖 3j 與該第五特徵圖 3i 相加，以產生並輸出一第五混合特徵圖 3k 至第四個該第一 CSP 模組 113；且第三個該第二加法器 124 將第三個 CMR 模組 131 輸出的該第六強化特徵圖 4j 與該第六特徵圖 4i 相加，以產生並輸出一第六混合特徵圖 4k 至第四個該第二 CSP 模組 123。第四個該第一 CSP 模組 113 對該第五混合特徵圖 3k 進行特徵提取以產生並輸出一第七特徵圖 31 至該第二照度網路 18；且第四個該第二 CSP 模組 123 對該第六混合特徵圖 4k 進行特徵提取以產生並輸出一第八特徵圖 41 至該第二照度網路 18；而該第三骨幹層 13 的第三個該第三 CSP 模組 132 對該第三強化融合特徵圖 5f 再次進行特徵擷取而產生並輸出一第五融合特徵圖 5g 至該第一多尺度層 16。

【0065】 接著，如圖 3C 的步驟 S40，該第一多尺度層 16 對該第五融合特徵圖 5g 進行基於不同尺度的特徵擷取，以產生並輸出一第一最終融合特徵圖 5h 至該全域融合偵測層 17。

【0066】 且如圖 3C 的步驟 S41，該全域融合偵測層 17 根據該第一最終融合特徵圖 5h 中的影像特徵進行候選框偵測及物件辨識，以產生並輸出具有複數個第一候選物件資訊的一第一候選特徵圖 171 至該基於錨框對齊模組 22。其中各該第一候選物件資訊至少包含一第一候選物件框及其對應的一信心指數(分數或機率)。

【0067】 且如圖 3C 的步驟 S42，該第二照度網路 18 將輸入的該第七特徵圖 31 以該第一權重加權以產生一加權後第三特徵圖 3m，且將該第八特徵圖 41 以該第二權重加權以產生一加權後第四特徵圖 4m，並將該加權後第三特徵圖 3m 和該加權後第四特徵圖 4m 輸出至該第二基於像素對齊模組 19。

【0068】 接著如圖 3C 的步驟 S43，該第二基於像素對齊模組 19 如同該第一基於像素對齊(校正)模組 15，根據輸入的該加權後第三特徵圖 3m 的特徵像素和該加權後第四特徵圖 4m 的特徵像素各自與一偏移場域之間的一偏移量，校正(位移)該加權後第三特徵圖 3m 的特徵像素和該加權後第四特徵圖 4m 的特徵像素的位置，而產生特徵像素對齊的一校正且加權後第三特徵圖 3n 和一校正且加權後第四特徵圖 4n，並將該校正且加權後第三特徵圖 3n 與該校

正且加權後第四特徵圖 4n 相疊合而產生並輸出一第五融合特徵圖 6a 至該第二多尺度層 20。

【0069】 接著如圖 3C 的步驟 S44，該第二多尺度層 20 對該第五融合特徵圖 6a 進行基於不同尺度的特徵擷取，以產生並輸出一第二最終融合特徵圖 6b 至該區域融合偵測層 21。

【0070】 接著如圖 3C 的步驟 S45，該區域融合偵測層 21 根據該第二最終融合特徵圖 6b 中的影像特徵進行候選框偵測及物件辨識，以產生並輸出具有複數個第二候選物件資訊的一第二候選特徵圖 211 至該基於錨框對齊模組 22。其中各該第二候選物件資訊 211 至少包含一第二候選物件框及其對應的一信心指數(分數或機率)。

【0071】 接著如圖 3D 的步驟 S46，該基於錨框對齊模組 22 根據該第一候選特徵圖 171 包含的該等第一候選物件框和該第二候選特徵圖 211 包含的該等第二候選物件框的相對偏移位置，校正(位移)該等第一候選物件框和該等第二候選物件框，以使校正後的該等第一候選物件框和校正後的該等第二候選物件框及其所涵蓋的影像物件能夠對齊，且據此產生並輸出具有複數個最終候選物件資訊的一最終特徵圖 22a 至該物件判定模組 23；其中各該最終候選物件資訊包含複數個最終候選物件框及其對應的一信心指數(分數或機率)。

【0072】 具體而言，如圖 5 所示，該基於錨框對齊模組 22 包含

依序串連的一特徵融合層 221、一錨框位移層 222 及一卷積層 223，該特徵融合層 221 主要將該第一候選特徵圖 171 和第二候選特徵圖 211 相疊合成為一融合特徵圖，再由該錨框位移層 222 根據融合特徵圖中該等第一候選物件框和該等第二候選物件框的相對偏移位置，校正該等第一候選物件框和該等第二候選物件框，使校正後的該等第一候選物件框和校正後的該等第二候選物件框對齊的同時，其所涵蓋的影像物件也能夠對齊，再將該融合特徵圖輸出至卷積層 223 進行卷積運算以降維，將特徵維度重塑，而產生具有該等最終候選物件資訊的該最終特徵圖 22a。與該基於錨框對齊模組 22 類似概念的相關技術可以參見「Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection」該篇論文中第 4 章「The Proposed Approach」其中 Alignment Process 的做法。

【0073】 最後如圖 3D 的步驟 S47，該物件判定模組 23 根據 DIOU-NMS 演算法之原理，從該最終特徵圖 22a 的該等最終候選物件資訊中選出最佳 (信心指數最高)的最終候選物件資訊，並將選出的最佳候選物件資訊(包含物件框及其對應的信心指數)標註於該可見光影像 3’及相對應的該熱影像 4’，因此在該可見光影像 3’及相對應的該熱影像 4’中會顯示框選物件的物件框；當然也可以只在該可見光影像 3’標註選出的最佳候選物件資訊。

【0074】 且藉由依序將每一組訓練用影像輸入該影像物件辨識模型 100，使影像物件辨識模型 100 重覆執行上述步驟 S31~S47，以反覆進行影像特徵擷取的訓練和深度學習，將使該影像物件辨識模型 100 的辨識率逐漸提升並收斂至一目標值，而獲得完成訓練的該影像物件辨識模型 100。

【0075】 因此，當該影像物件辨識模型 100 被訓練完成後，將一組待辨識影像中的一待辨識熱影像和一待辨識可見光影像被輸入該影像物件辨識模型 100 時，該影像物件辨識模型 100 即可辨識出該待辨識可見光影像中的物件，並於輸出的該待辨識可見光影像中，將辨識出來的物件以物件框框選並標註其類別(例如人、車(汽車、卡車、機車、公車等)、動物(狗、貓、馬等)、植物等)。值得一提的是，本實施例也可應用但不限於台灣第 110104936 號專利申請案提供的雙影像融合方法，將該待辨識熱影像和該待辨識可見光影像融合成一融合影像後輸出，並根據影像辨識結果，將該融合影像中被辨識的物件框選並標註其類別。

【0076】 再參見圖 6 所示，是本發明的第二實施例，其與第一實施例不同處在於該影像物件辨識模型的該第一骨幹層 11 只採用二個第一 CSP 模組 113，該第二骨幹層 12 也只採用二個第二 CSP 模組 123，該影像物件辨識模型只採用一個 CMR 模組 131，且該第三骨幹層 13 只採用一個第三 CSP 模組 132；因此 CMR 模組

131 是直接將產生的一強化融合特徵圖 5f 輸出至第三 CSP 模組 132，第三 CSP 模組 132 對強化融合特徵圖 5f 進行特徵提取以產生並輸出一特徵圖至該第一多尺度層 16；且第二個該第一 CSP 模組 113 對加法器 114 提供的一混合特徵圖 3k 進行特徵提取所產生的一特徵圖 31 直接輸出至第二照度網路 18，第二個該第二 CSP 模組 123 對加法器 124 提供的一混合特徵圖 4k 進行特徵提取所產生的一特徵圖 41 直接輸出至第二照度網路 18。

【0077】 再參見圖 7 所示，是本發明的第三實施例，其與第一實施例不同處在於該影像物件辨識模型 200 只採用一個 CMR 模組 131 且該第三骨幹層 13 只採用一個 CSP 模組 132，且該 CMR 模組 131 是根據該第一骨幹層 11 的倒數第二個 CSP 層 113 和該第二骨幹層 11 的倒數第二個 CSP 層 123 輸出的特徵圖及該第一基於像素對齊模組 15 輸出的特徵圖產生相對應的三個強化特徵圖並分別輸出至該第一骨幹層 11 的加法器 114、該第二骨幹層 12 的加法器 124 和該第三骨幹層 13 該 CSP 模組 132。

【0078】 綜上所述，上述實施例藉由照度網路根據可見光影像呈現的照度調整可見光影像的特徵圖和熱影像的特徵圖的比重，並藉由基於像素對齊模組和基於錨框對齊模組改善雙影像因視野不同造成之特徵偏移，因此對於雙影像之間的特徵偏移和視野差異的容忍度高；且藉由跨模態特徵強化模組針對雙影像特徵的不平衡，

套用自注意力機制強化不同模態的特徵特點，改善模型經過訓練後的特徵丟失，並融合全域(或全局)(Global)特徵和區域(或局部)(Local)特徵，而達到多物件偵測效果的提升。且本實施例的影像物件辨識模型100藉由對在同一時間拍攝的可見光影像及熱影像進行影像物件辨識，可同時取得這兩種影像的特徵，而利用影像特徵互補的效果，進行全天候的影像辨識並提升物件辨識率，使影像物件辨識不致受限於時段、天候或環境的變化，也不需根據時段、天候或環境變化不斷地切換不同的影像辨識機制，確實達到本發明的功效與目的。

【0079】 惟以上所述者，僅為本發明之實施例而已，當不能以此限定本發明實施之範圍，凡是依本發明申請專利範圍及專利說明書內容所作之簡單的等效變化與修飾，皆仍屬本發明專利涵蓋之範圍內。

【符號說明】

【0080】

100、200、300：影像物件辨識模型

11：第一骨幹層

111：聚焦層

112：第一跨階段局部網路(第一 CSPNet)

113：第一跨階段局部模組(第一 CSP 模組)

114：加法器

12：第二骨幹層

121：聚焦層

122：第二跨階段局部網路(第二 CSPNet)

123：第二 CSP 模組

124：加法器

13：第三骨幹層

131：跨模態特徵強化(CMR)模組

132：第三跨階段局部模組(第三 CSP 模組)

133：加法器

134：第一特徵強化層

135：第二特徵強化層

136：第三特徵強化層

137：第四特徵強化層

14：第一照度網路

15：第一基於像素對齊模組

16：第一多尺度層

17：全域融合偵測層

171：第一候選特徵圖

18：第二照度網路

19：第二基於像素對齊模組

20：第二多尺度層

21：區域融合偵測層

211：第二候選特徵圖

22：基於錨框對齊模組

22a：最終特徵圖

221：特徵融合層

222：錨框位移層

223：卷積層

23：物件判定模組

3、3’：可見光影像

4、4’：熱影像

3a：第一特徵圖

3b：加權後第一特徵圖

3c：校正且加權後第一特徵圖

3d：第一強化特徵圖

3e：第一混合特徵圖

3f：第三特徵圖

3g：第三強化特徵圖

3h：第三混合特徵圖

3i：第五特徵圖

3j：第五強化特徵圖

3k：第五混合特徵圖、混合特徵圖

3l：第七特徵圖、特徵圖

3m：加權後第三特徵圖

3n：校正且加權後第三特徵圖

4a：第二特徵圖

4b：加權後第二特徵圖

4c：校正且加權後第二特徵圖

4d：第二強化特徵圖

4e：第二混合特徵圖

4f：第四特徵圖

4g：第四強化特徵圖

4h：第四混合特徵圖

4i：第六特徵圖

4j：第六強化特徵圖

4k：第六混合特徵圖、混合特徵圖

4l：第八特徵圖、特徵圖

4m：加權後第四特徵圖

4n：校正且加權後第四特徵圖

5a：第一融合特徵圖

5b：第一強化融合特徵圖

5c：第三融合特徵圖

5d：第二強化融合特徵圖

5e：第四融合特徵圖

5f：第三強化融合特徵圖、強化融合特徵圖

5g：第五融合特徵圖

5h：第一最終融合特徵圖

6a：第五融合特徵圖

6b：第二最終融合特徵圖

S1~S3、S31~S47：步驟

【發明申請專利範圍】

【請求項1】一種影像物件辨識模型的訓練方法，包括：

A、一電腦裝置的一處理單元執行預先載入之待訓練的一影像物件辨識模型，該影像物件辨識模型包括一第一骨幹層、一第二骨幹層、一第三骨幹層、與該第三骨幹層串接的一跨模態特徵強化模組、與該第一骨幹層和該第二骨幹層連接的一第一照度網路、與該第一照度網路和該跨模態特徵強化模組連接的一第一基於像素對齊模組、與該第三骨幹層連接的一第一多尺度層、與該第一多尺度層連接的一全域融合偵測層、與該第一骨幹層和該第二骨幹層連接的一第二照度網路、與該第二照度網路連接的一第二基於像素對齊模組、與該第二基於像素對齊模組連接的一第二多尺度層、與該第二多尺度層連接的一區域融合偵測層、與該全域融合偵測層和該區域融合偵測層連接的一基於錨框對齊模組以及與該基於錨框對齊模組連接的一物件判定模組；及

B、該電腦裝置提供複數組訓練用影像給該處理單元，每一組訓練用影像包含在同一時間拍攝且內容重疊的一可見光影像與一熱影像；該處理單元依序將每一組訓練用影像輸入該影像物件辨識模型，以藉由反覆執行下述動作訓練該影像物件辨識模型：

B1、將該可見光影像和該熱影像各別對應輸入該第一骨幹層和該第二骨幹層，並將該可見光影像輸入該第一照度網路和該第二照度網路，使該第一骨幹層對該

可見光影像進行區域特徵提取並輸出一第一特徵圖至該第一照度網路和該跨模態特徵強化模組，並使該第二骨幹層對該熱影像進行區域特徵提取並輸出一第二特徵圖至該第一照度網路和該跨模態特徵強化模組；且該第一照度網路和該第二照度網路偵測該可見光影像中的照度以輸出與明亮環境相關且代表明亮的機率的一第一權重和與陰暗環境相關且代表陰暗的機率的一第二權重；

B2、該第一照度網路將該第一特徵圖以該第一權重加權以產生一加權後第一特徵圖，且將該第二特徵圖以該第二權重加權以產生一加權後第二特徵圖，並將該加權後第一特徵圖和該加權後第二特徵圖輸出至該第一基於像素對齊模組；

B3、該第一基於像素對齊模組根據該加權後第一特徵圖的特徵像素和該加權後第二特徵圖的特徵像素各自與一偏移場域之間的一偏移量，校正該加權後第一特徵圖和該加權後第二特徵圖，以產生特徵像素對齊的一校正且加權後第一特徵圖和一校正且加權後第二特徵圖，並將該校正且加權後第一特徵圖與該校正且加權後第二特徵圖相疊合而產生並輸出一第一融合特徵圖至該跨模態特徵強化模組；

B4、該跨模態特徵強化模組利用基於移位窗口的自注意力機制對輸入的該第一特徵圖、該第二特徵圖和該第一融合特徵圖進行全域特徵擷取，以對應產生三個

強化特徵圖，並分別輸出三個強化特徵圖至相對應的該第一骨幹層、該第二骨幹層和該第三骨幹層；

B5、該第三骨幹層對該跨模態特徵強化模組輸出的該強化特徵圖進行特徵擷取而產生並輸出一特徵圖至該第一多尺度層；該第一骨幹層基於該第一特徵圖與該跨模態特徵強化模組輸出的該強化特徵圖進行特徵提取以產生並輸出一特徵圖至該第二照度網路，且該第二骨幹層基於該第二特徵圖與該跨模態特徵強化模組輸出的該強化特徵圖進行特徵提取以產生並輸出一特徵圖至該第二照度網路；

B6、該第一多尺度層對該第三骨幹層輸出的該特徵圖進行基於不同尺度的特徵擷取，以產生並輸出一特徵圖至該全域融合偵測層；該全域融合偵測層根據該第一多尺度層輸出的該特徵圖中的影像特徵進行候選框偵測和物件辨識，以產生並輸出具有複數個第一候選物件資訊的一特徵圖至該基於錨框對齊模組；

B7、該第二照度網路將該第一骨幹層輸出的該特徵圖以該第一權重加權以產生一加權後特徵圖，且將該第二骨幹層輸出的該特徵圖以該第二權重加權以產生一加權後特徵圖，並將該二個加權後特徵圖輸出至該第二基於像素對齊模組；

B8、該第二基於像素對齊模組根據該第二照度網路輸出的該二個加權後特徵圖的特徵像素各自與一偏移場域之間的一偏移量，校正該二個加權後特徵圖，以

產生特徵像素對齊的二個校正且加權後特徵圖，並將該二個校正且加權後特徵圖相疊合而產生並輸出一融合特徵圖至該第二多尺度層；該第二多尺度層對該第二基於像素對齊模組輸出的該融合特徵圖進行基於不同尺度的特徵擷取，以產生並輸出一特徵圖至該區域融合偵測層；該區域融合偵測層根據該第二多尺度層輸出的該特徵圖中的影像特徵進行候選框偵測和物件辨識，以產生並輸出具有複數個第二候選物件框的一特徵圖至該基於錨框對齊模組；

B9、該基於錨框對齊模組根據輸入的二個特徵圖對應的該等第一候選物件框和該等第二候選物件框的相對偏移位置，校正該等第一候選物件框和該等第二候選物件框，使校正後的該等第一候選物件框和校正後的該等第二候選物件框及其所涵蓋的影像物件對齊，以此產生並輸出具有複數個最終候選物件框的一特徵圖至該物件判定模組；及

B10、該物件判定模組根據該基於錨框對齊模組輸出的該特徵圖中的該等最終候選物件框所對應的信心指數，從該等最終候選物件框選出最佳的候選物件框，並將選出的最佳候選物件框顯示在該可見光影像中。

【請求項2】如請求項1所述影像物件辨識模型的訓練方法，其中每一組訓練用影像包含的該熱影像是預先以一座標投影矩陣校正的校正後熱影像，使該校正後熱影像能與相對應的

該可見光影像對齊，且該座標投影矩陣是根據該熱影像與相對應的該可見光影像兩者的視野差異而求得。

【請求項3】如請求項1所述影像物件辨識模型的訓練方法，其中該跨模態特徵強化模組包含依序串連的四個特徵強化層，該等特徵強化層包含用於正規化輸入的特徵的一正規化層；第一個特徵強化層包含的W-MSA層和第三個特徵強化層包含的SW-MSA層對輸入的特徵進行多頭自注意力處理以提取特徵；第二個和第四個特徵強化層包含的一卷積層對輸入的特徵進行卷積運算以提取特徵。

【請求項4】如請求項1所述影像物件辨識模型的訓練方法，其中該基於錨框對齊模組包含依序串連的一特徵融合層、一錨框位移層及一卷積層，該特徵融合層將輸入的二個特徵圖相疊合成為一融合特徵圖，再由該錨框位移層根據該融合特徵圖中該等第一候選物件框和該等第二候選物件框的相對偏移位置，校正該等第一候選物件框和該等第二候選物件框，使校正後的該等第一候選物件框和校正後的該等第二候選物件框及其所涵蓋的影像物件對齊，再由該卷積層對該融合特徵圖進行卷積運算以降維，並輸出具有該等最終候選物件框的該特徵圖。

【請求項5】如請求項1所述影像物件辨識模型的訓練方法，其中該跨模態特徵強化模組是第一個跨模態特徵強化模組，且該影像物件辨識模型還包括與第一個跨模態特徵強化模組和該第三骨幹層串接的第二個跨模態特徵強化模組和第三個跨模態特徵強化模組；且步驟B5還包括下列步驟：

B51、該第一骨幹層基於該第一特徵圖與該強化特徵圖進行特徵提取而產生的該特徵圖被輸入第二個跨模態特徵強化模組，且該第二骨幹層基於該第二特徵圖與該強化特徵圖進行特徵提取而產生的該特徵圖被輸入第二個跨模態特徵強化模組，該第三骨幹層對該強化特徵圖進行特徵擷取而產生的該特徵圖被輸入第二個跨模態特徵強化模組；

B52、第二個跨模態特徵強化模組對輸入的三個特徵圖進行特徵強化，以對應產生三個強化特徵圖，並分別輸出該三個強化特徵圖至相對應的該第一骨幹層、該第二骨幹層和該第三骨幹層；

B53、該第一骨幹層基於輸入該第二個跨模態特徵強化模組的該特徵圖和該第二個跨模態特徵強化模組輸出的強化特徵圖進行特徵提取而產生並輸出一特徵圖至該第三個跨模態特徵強化模組；該第二骨幹層基於輸入該第二個跨模態特徵強化模組的該特徵圖和該第二個跨模態特徵強化模組輸出的強化特徵圖進行特徵提取而產生並輸出一特徵圖至該第三個跨模態特徵強化模組；該第三骨幹層對第二個跨模態特徵強化模組輸出的強化特徵圖進行特徵提取而產生並輸出一特徵圖至該第三個跨模態特徵強化模組；

B54、第三個跨模態特徵強化模組對輸入的三個特徵圖進行特徵強化，以對應產生三個強化特徵圖，並分別輸

出三個強化特徵圖至相對應的該第一骨幹層、該第二骨幹層和該第三骨幹層；及

B55、該第三骨幹層對第三個跨模態特徵強化模組輸出的強化特徵圖進行特徵提取而產生並輸出該特徵圖至該第一多尺度層；該第一骨幹層基於輸入該第三個跨模態特徵強化模組的該特徵圖和該第三個跨模態特徵強化模組輸出的強化特徵圖進行特徵提取而產生並輸出該特徵圖至該第二照度網路；該第二骨幹層基於輸入該第三個跨模態特徵強化模組的該特徵圖和該第三個跨模態特徵強化模組輸出的強化特徵圖進行特徵提取而產生並輸出該特徵圖至該第二照度網路。

【請求項6】一種影像物件辨識模型，其係根據請求項1至5其中任一項所述影像物件辨識模型的訓練方法訓練而成，而能接受包含在同一時間拍攝且內容重疊的一待辨識可見光影像與一待辨識熱影像的一組待辨識影像，以根據該待辨識可見光影像與該待辨識熱影像辨識出該待辨識可見光影像中的物件。

【請求項7】一種電腦可存取的記錄媒體，其中存有如請求項1至5其中任一項所述影像物件辨識模型的訓練方法其中所述的該影像物件辨識模型，且該影像物件辨識模型藉由如請求項1至5其中任一項所述影像物件辨識模型的訓練方法訓練完成後，能根據輸入之在同一時間拍攝且內容重疊的一可見光影像與一熱影像辨識該可見光影像中的物件。

【發明圖式】

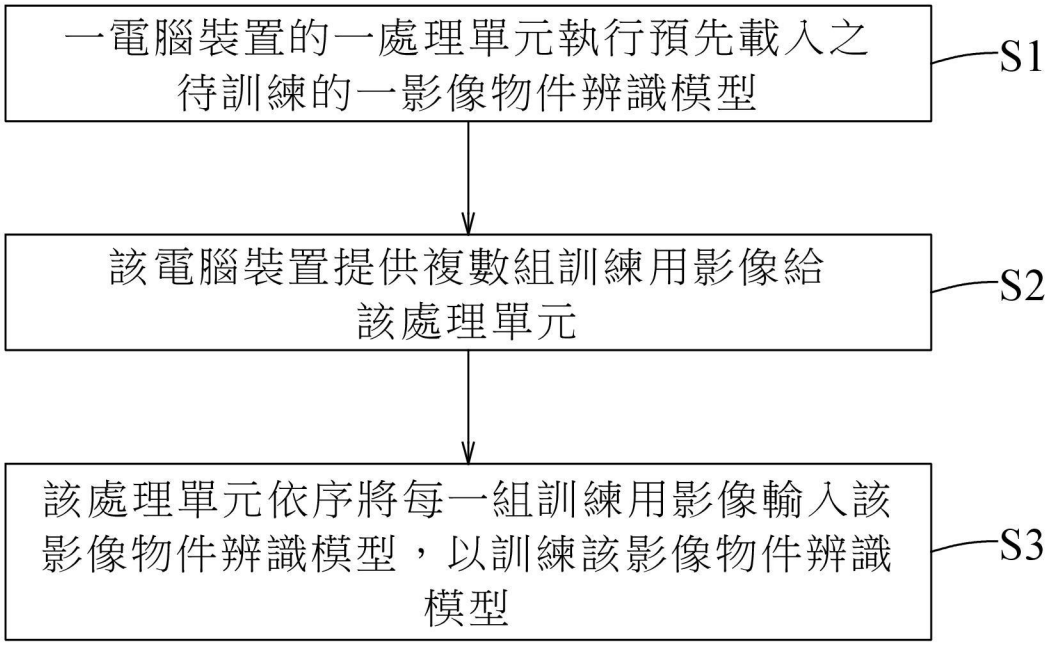
一電腦裝置的一處理單元執行預先載入之待訓練的一影像物件辨識模型
S1
該電腦裝置提供複數組訓練用影像給該處理單元
S2
該處理單元依序將每一組訓練用影像輸入該影像物件辨識模型，以訓練該影像物件辨識模型
S3

圖 1

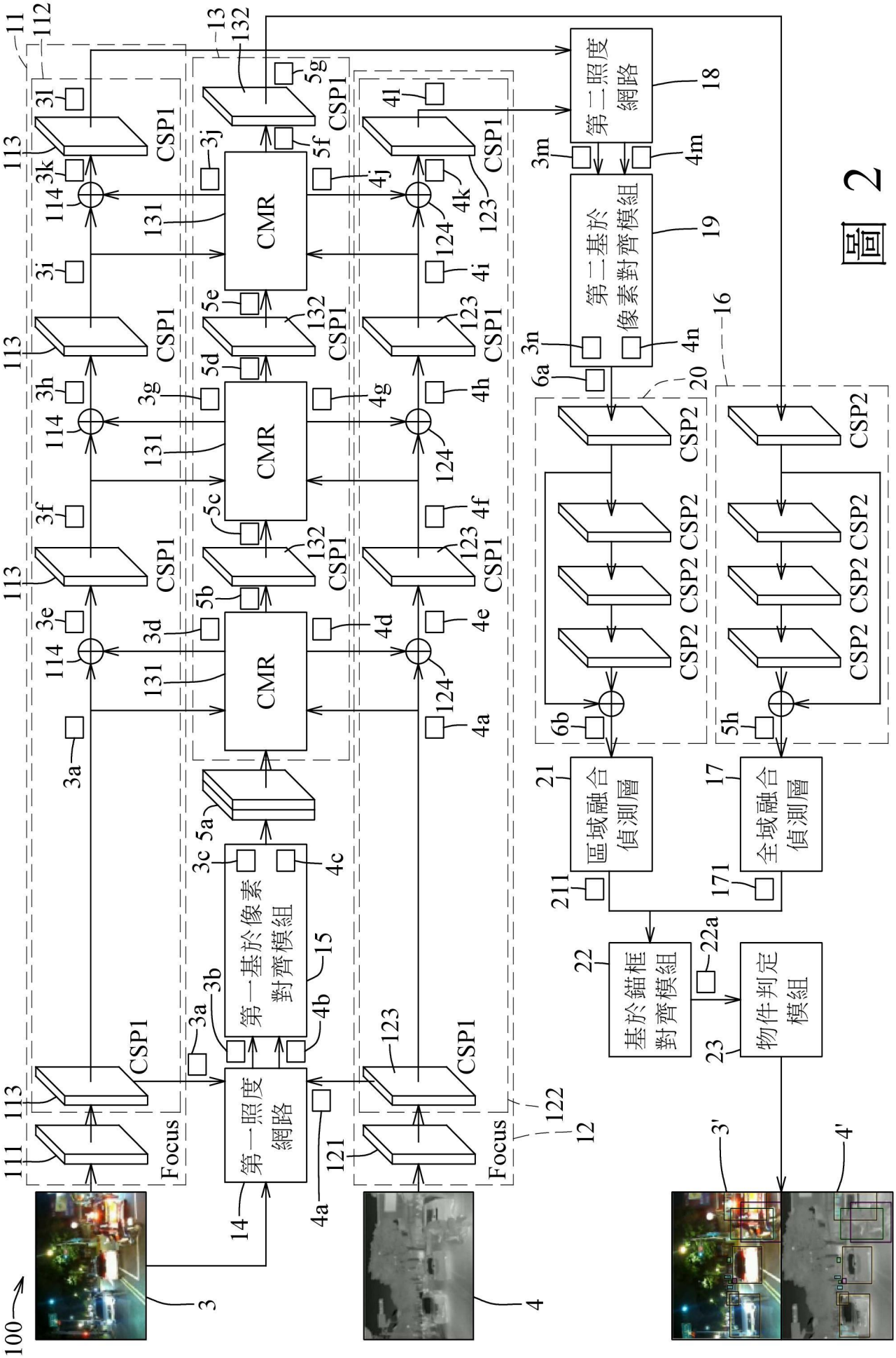
100
3
4
Focus
CSP1
CMR
第一照度網路
第一基於像素對齊模組
第二照度網路
第二基於像素對齊模組
區域融合偵測層
全域融合偵測層
基於錨框對齊模組
物件判定模組
CSP2
3'
4'

圖 2

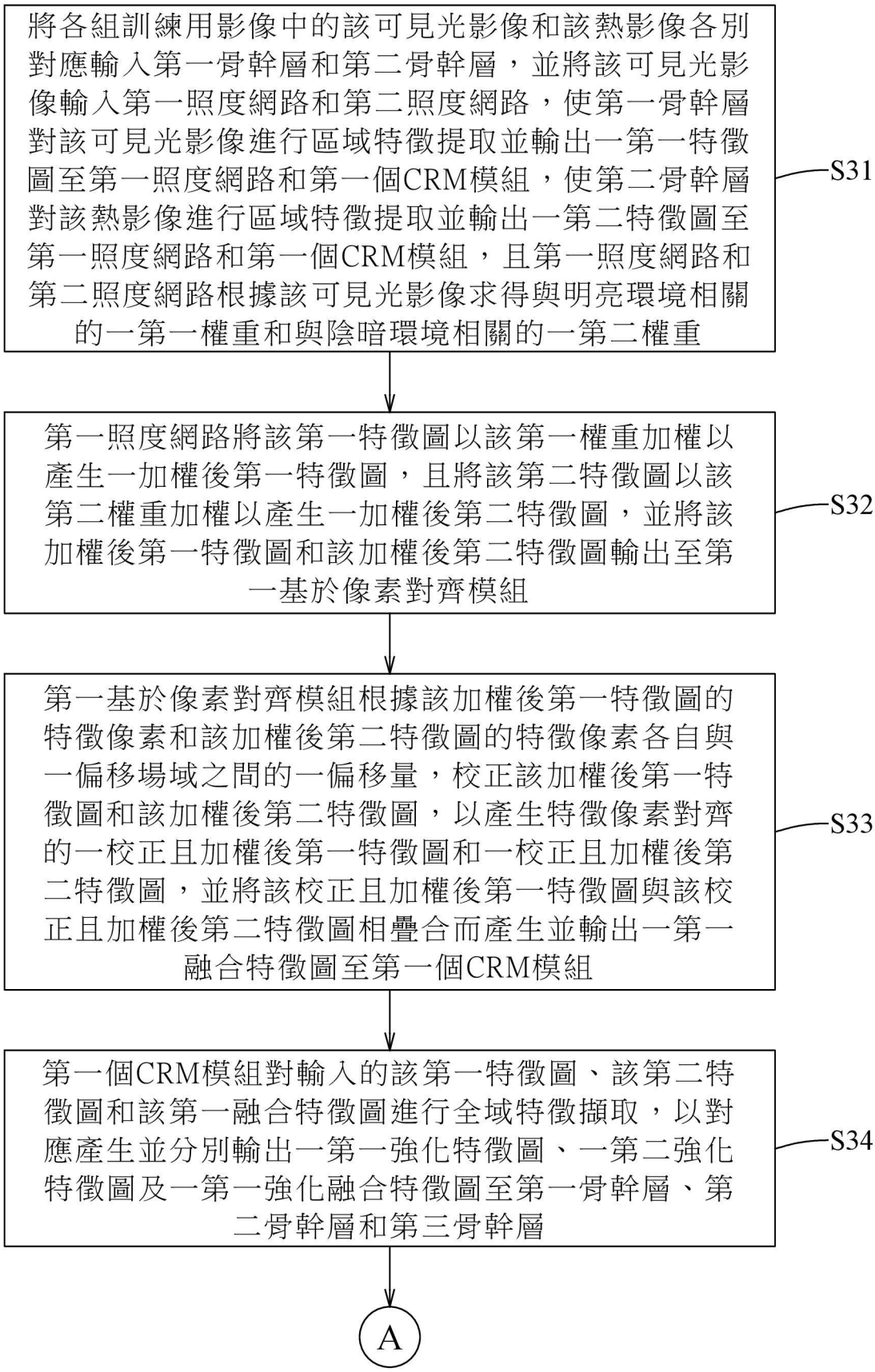
將各組訓練用影像中的該可見光影像和該熱影像各別對應輸入第一骨幹層和第二骨幹層，並將該可見光影像輸入第一照度網路和第二照度網路，使第一骨幹層對該可見光影像進行區域特徵提取並輸出一第一特徵圖至第一照度網路和第一個CRM模組，使第二骨幹層對該熱影像進行區域特徵提取並輸出一第二特徵圖至第一照度網路和第一個CRM模組，且第一照度網路和第二照度網路根據該可見光影像求得與明亮環境相關的一第一權重和與陰暗環境相關的一第二權重
S31
第一照度網路將該第一特徵圖以該第一權重加權以產生一加權後第一特徵圖，且將該第二特徵圖以該第二權重加權以產生一加權後第二特徵圖，並將該加權後第一特徵圖和該加權後第二特徵圖輸出至第一基於像素對齊模組
S32
第一基於像素對齊模組根據該加權後第一特徵圖的特徵像素和該加權後第二特徵圖的特徵像素各自與一偏移場域之間的一偏移量，校正該加權後第一特徵圖和該加權後第二特徵圖，以產生特徵像素對齊的一校正且加權後第一特徵圖和一校正且加權後第二特徵圖，並將該校正且加權後第一特徵圖與該校正且加權後第二特徵圖相疊合而產生並輸出一第一融合特徵圖至第一個CRM模組
S33
第一個CRM模組對輸入的該第一特徵圖、該第二特徵圖和該第一融合特徵圖進行全域特徵擷取，以對應產生並分別輸出一第一強化特徵圖、一第二強化特徵圖及一第一強化融合特徵圖至第一骨幹層、第二骨幹層和第三骨幹層
S34
A

圖 3A

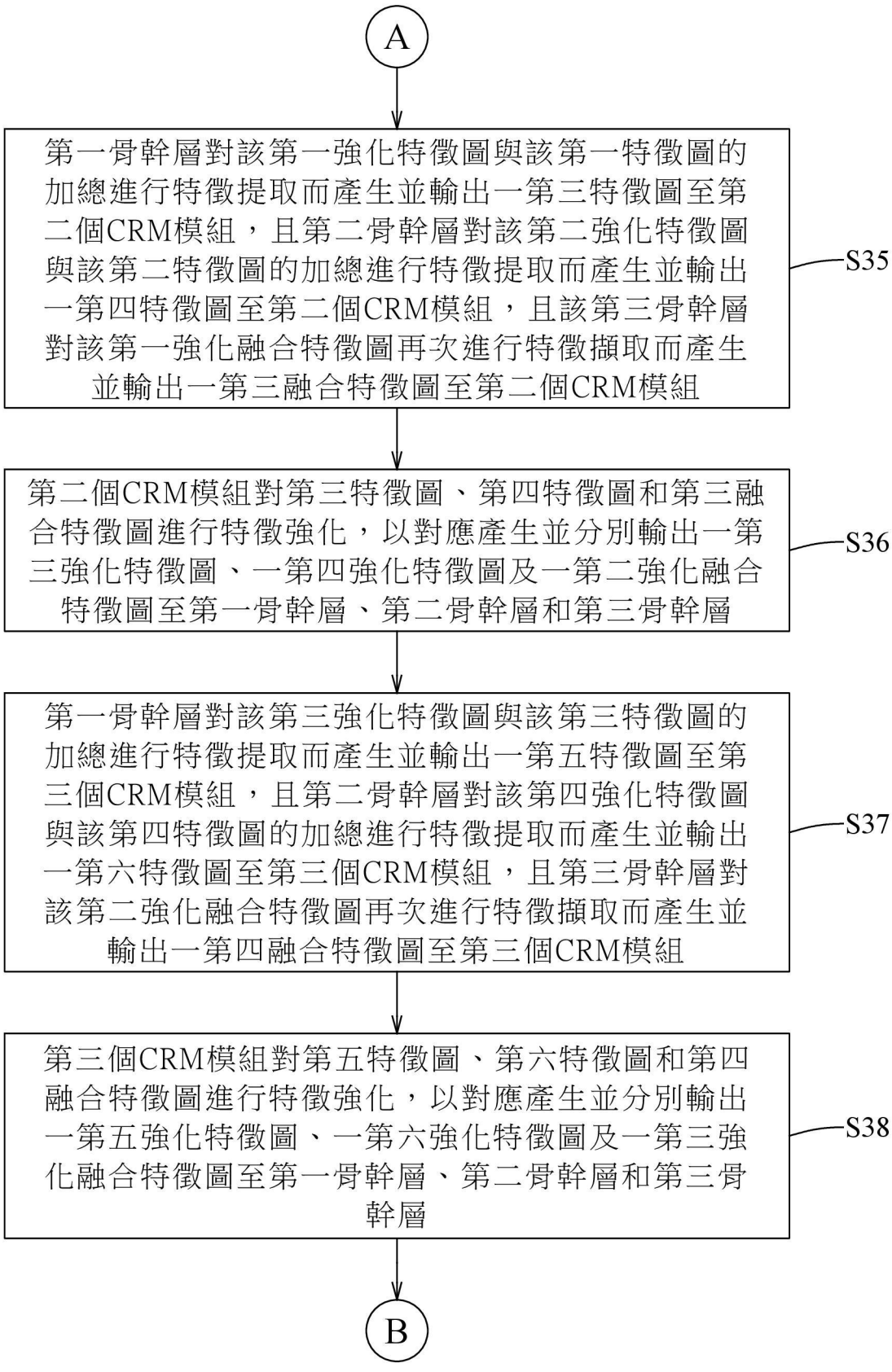
A
第一骨幹層對該第一強化特徵圖與該第一特徵圖的加總進行特徵提取而產生並輸出一第三特徵圖至第二個CRM模組，且第二骨幹層對該第二強化特徵圖與該第二特徵圖的加總進行特徵提取而產生並輸出一第四特徵圖至第二個CRM模組，且該第三骨幹層對該第一強化融合特徵圖再次進行特徵擷取而產生並輸出一第三融合特徵圖至第二個CRM模組
S35
第二個CRM模組對第三特徵圖、第四特徵圖和第三融合特徵圖進行特徵強化，以對應產生並分別輸出一第三強化特徵圖、一第四強化特徵圖及一第二強化融合特徵圖至第一骨幹層、第二骨幹層和第三骨幹層
S36
第一骨幹層對該第三強化特徵圖與該第三特徵圖的加總進行特徵提取而產生並輸出一第五特徵圖至第三個CRM模組，且第二骨幹層對該第四強化特徵圖與該第四特徵圖的加總進行特徵提取而產生並輸出一第六特徵圖至第三個CRM模組，且第三骨幹層對該第二強化融合特徵圖再次進行特徵擷取而產生並輸出一第四融合特徵圖至第三個CRM模組
S37
第三個CRM模組對第五特徵圖、第六特徵圖和第四融合特徵圖進行特徵強化，以對應產生並分別輸出一第五強化特徵圖、一第六強化特徵圖及一第三強化融合特徵圖至第一骨幹層、第二骨幹層和第三骨幹層
S38
B

圖 3B

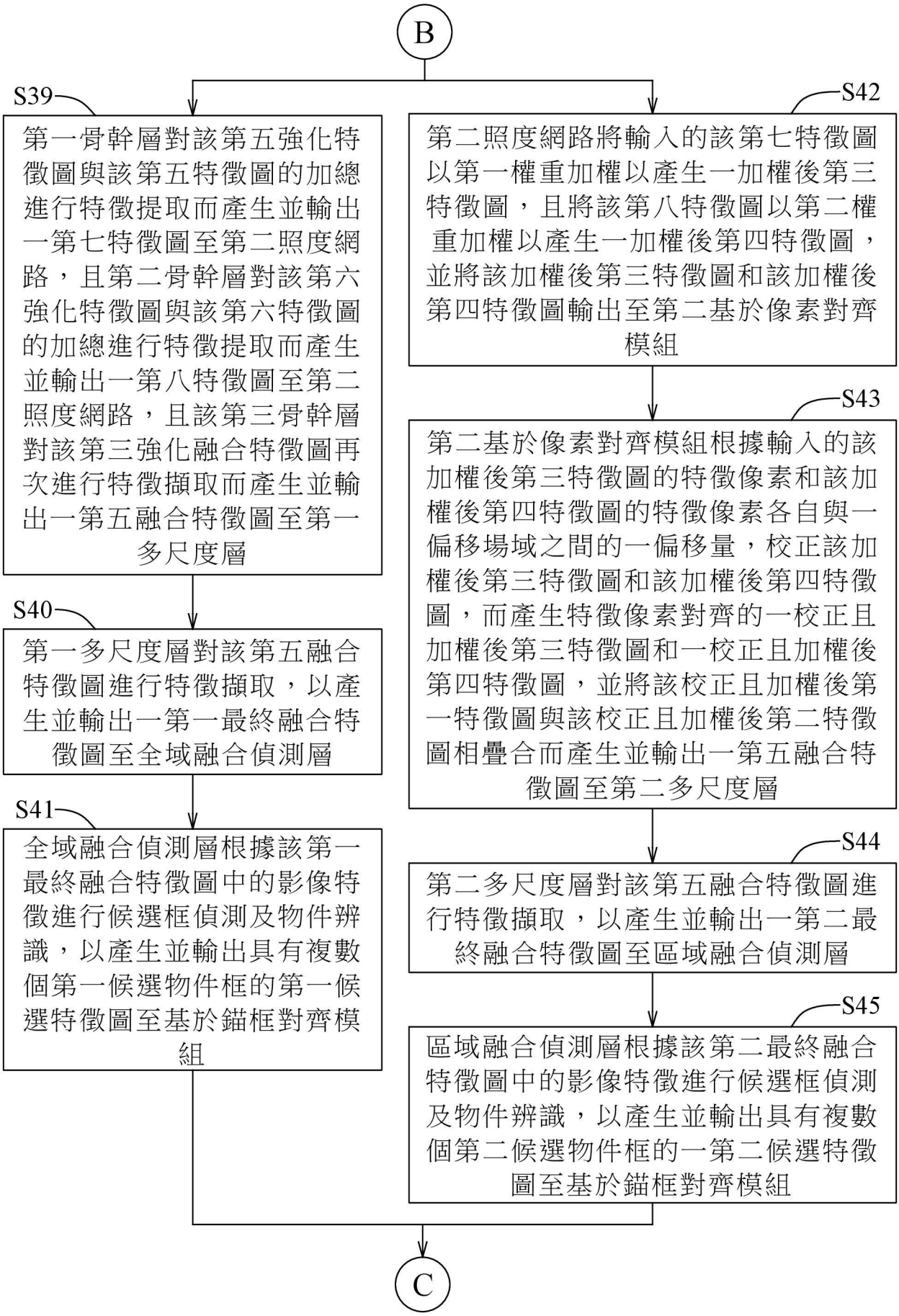
B
S39
第一骨幹層對該第五強化特徵圖與該第五特徵圖的加總進行特徵提取而產生並輸出一第七特徵圖至第二照度網路，且第二骨幹層對該第六強化特徵圖與該第六特徵圖的加總進行特徵提取而產生並輸出一第八特徵圖至第二照度網路，且該第三骨幹層對該第三強化融合特徵圖再次進行特徵擷取而產生並輸出一第五融合特徵圖至第一多尺度層
S40
第一多尺度層對該第五融合特徵圖進行特徵擷取，以產生並輸出一第一最終融合特徵圖至全域融合偵測層
S41
全域融合偵測層根據該第一最終融合特徵圖中的影像特徵進行候選框偵測及物件辨識，以產生並輸出具有複數個第一候選物件框的第一候選特徵圖至基於錨框對齊模組
S42
第二照度網路將輸入的該第七特徵圖以第一權重加權以產生一加權後第三特徵圖，且將該第八特徵圖以第二權重加權以產生一加權後第四特徵圖，並將該加權後第三特徵圖和該加權後第四特徵圖輸出至第二基於像素對齊模組
S43
第二基於像素對齊模組根據輸入的該加權後第三特徵圖的特徵像素和該加權後第四特徵圖的特徵像素各自與一偏移場域之間的一偏移量，校正該加權後第三特徵圖和該加權後第四特徵圖，而產生特徵像素對齊的一校正且加權後第三特徵圖和一校正且加權後第四特徵圖，並將該校正且加權後第一特徵圖與該校正且加權後第二特徵圖相疊合而產生並輸出一第五融合特徵圖至第二多尺度層
S44
第二多尺度層對該第五融合特徵圖進行特徵擷取，以產生並輸出一第二最終融合特徵圖至區域融合偵測層
S45
區域融合偵測層根據該第二最終融合特徵圖中的影像特徵進行候選框偵測及物件辨識，以產生並輸出具有複數個第二候選物件框的一第二候選特徵圖至基於錨框對齊模組
C

圖 3C

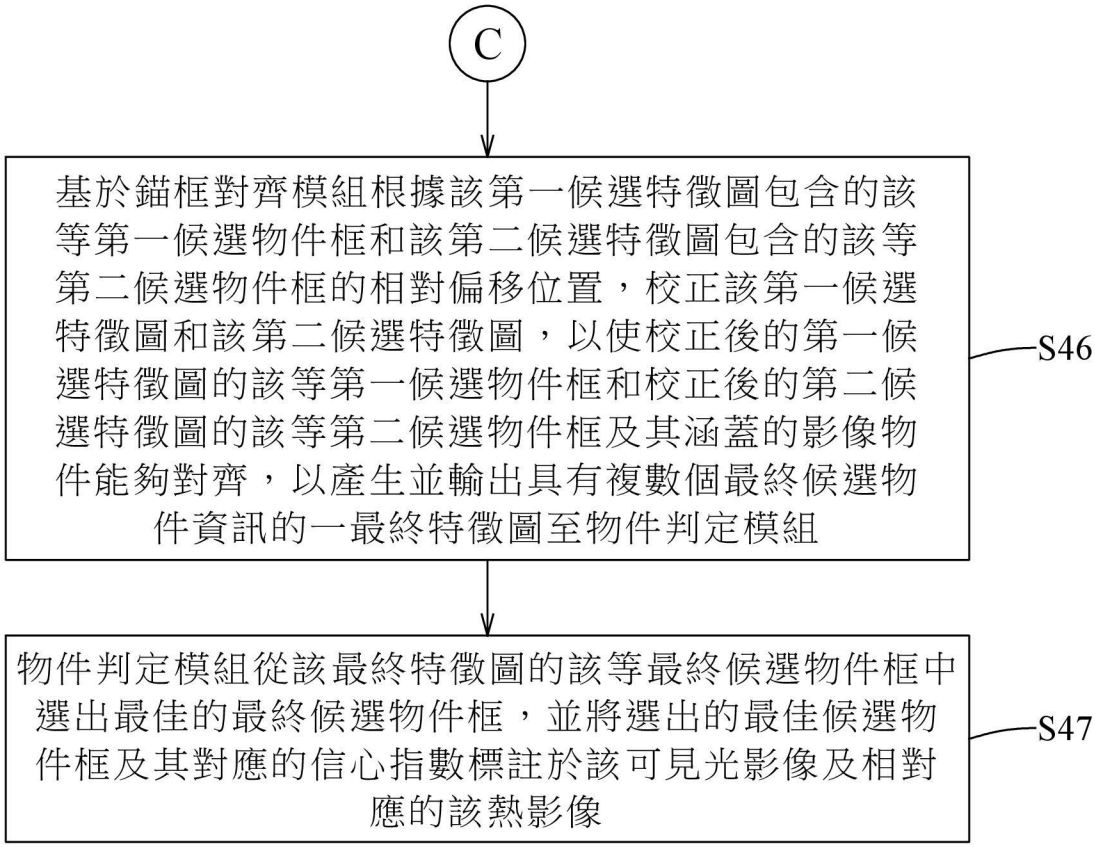
C
基於錨框對齊模組根據該第一候選特徵圖包含的該等第一候選物件框和該第二候選特徵圖包含的該等第二候選物件框的相對偏移位置，校正該第一候選特徵圖和該第二候選特徵圖，以使校正後的第一候選特徵圖的該等第一候選物件框和校正後的第二候選特徵圖的該等第二候選物件框及其涵蓋的影像物件能夠對齊，以產生並輸出具有複數個最終候選物件資訊的一最終特徵圖至物件判定模組
S46
物件判定模組從該最終特徵圖的該等最終候選物件框中選出最佳的最終候選物件框，並將選出的最佳候選物件框及其對應的信心指數標註於該可見光影像及相對應的該熱影像
S47

圖 3D

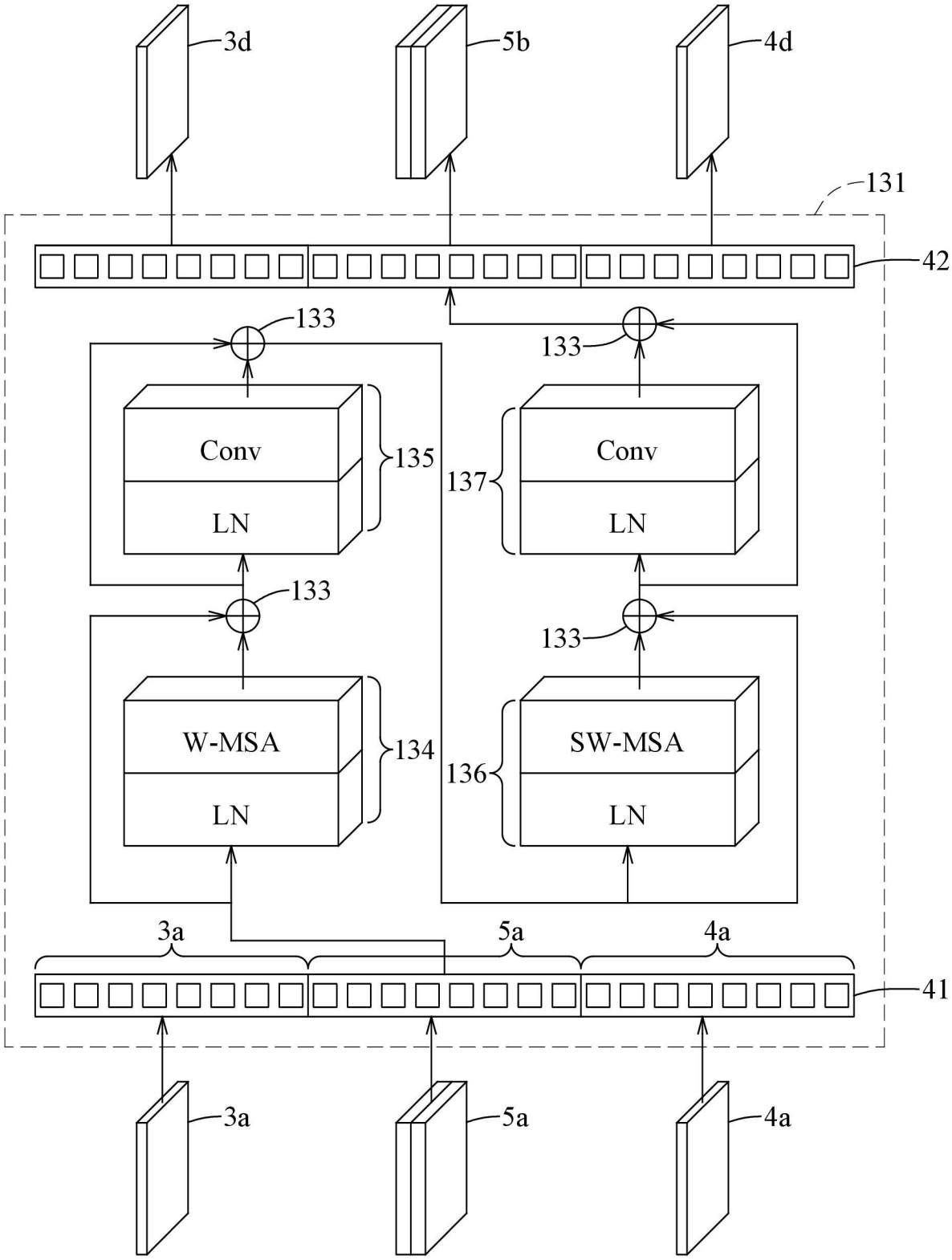
3d
5b
4d
131
42
133
133
Conv
LN
135
137
Conv
LN
133
133
W-MSA
LN
134
136
SW-MSA
LN
3a
5a
4a
41
3a
5a
4a

圖 4

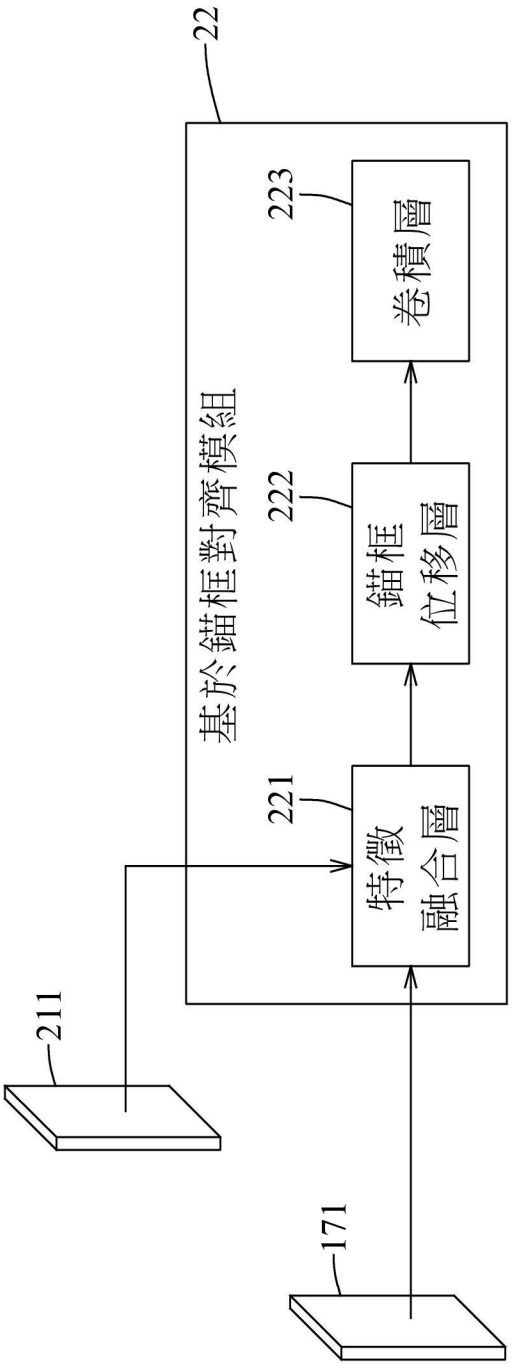

圖 5

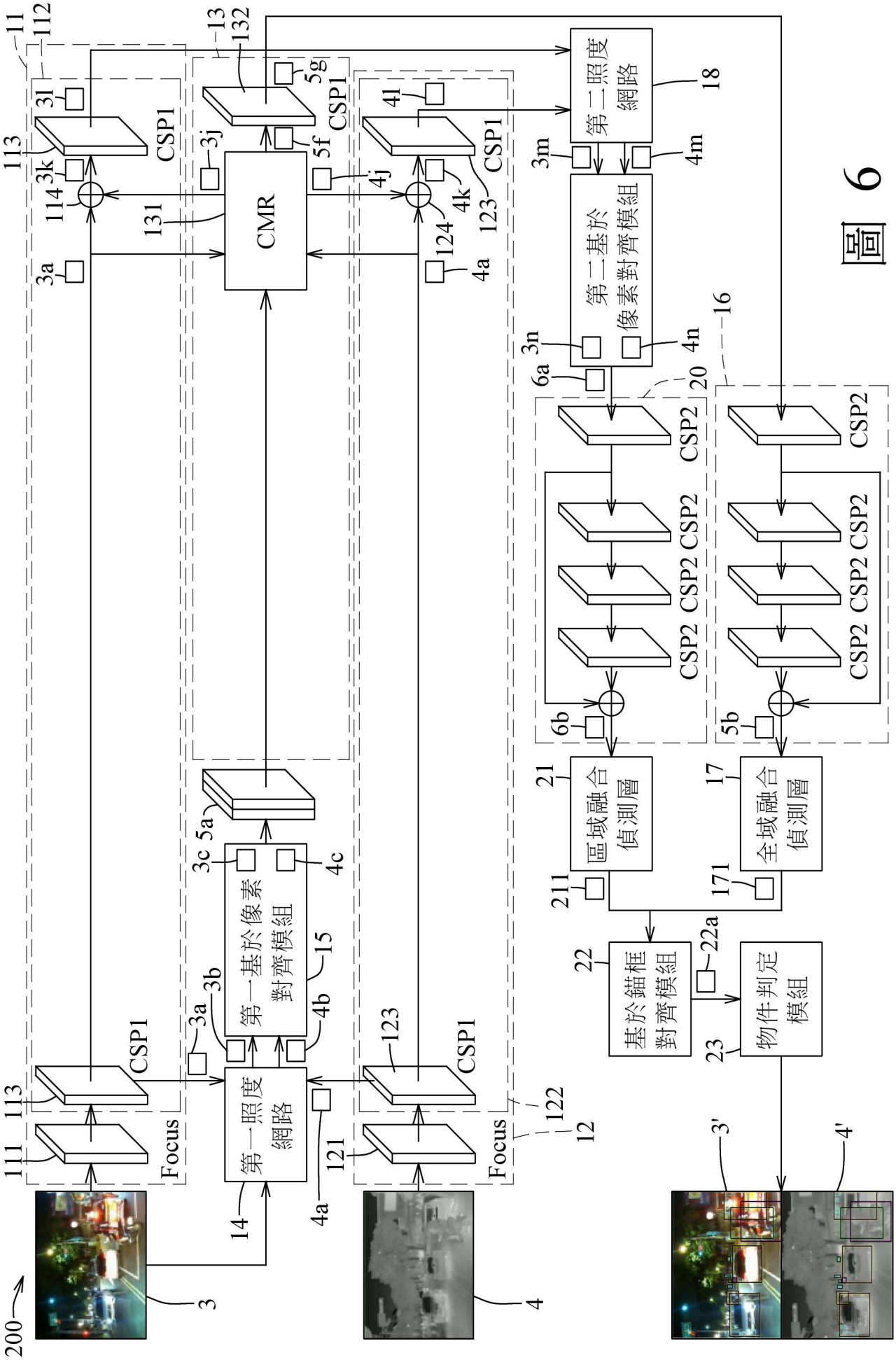
200
3
4
14
4a
121
12
122
Focus
111
113
CSP1
3a
3b
第一照度網路
第一基於像素對齊模組
15
4b
3c
5a
4c
CMR
131
132
13
112
11
3l
3k
114
3j
5f
5g
CSP1
41
4j
4k
123
124
4a
第二照度網路
18
3m
4m
第二基於像素對齊模組
16
3n
4n
6a
20
CSP2
6b
5b
21
17
區域融合偵測層
全域融合偵測層
211
171
22
22a
基於錨框對齊模組
物件判定模組
23
3'
4'

圖 6

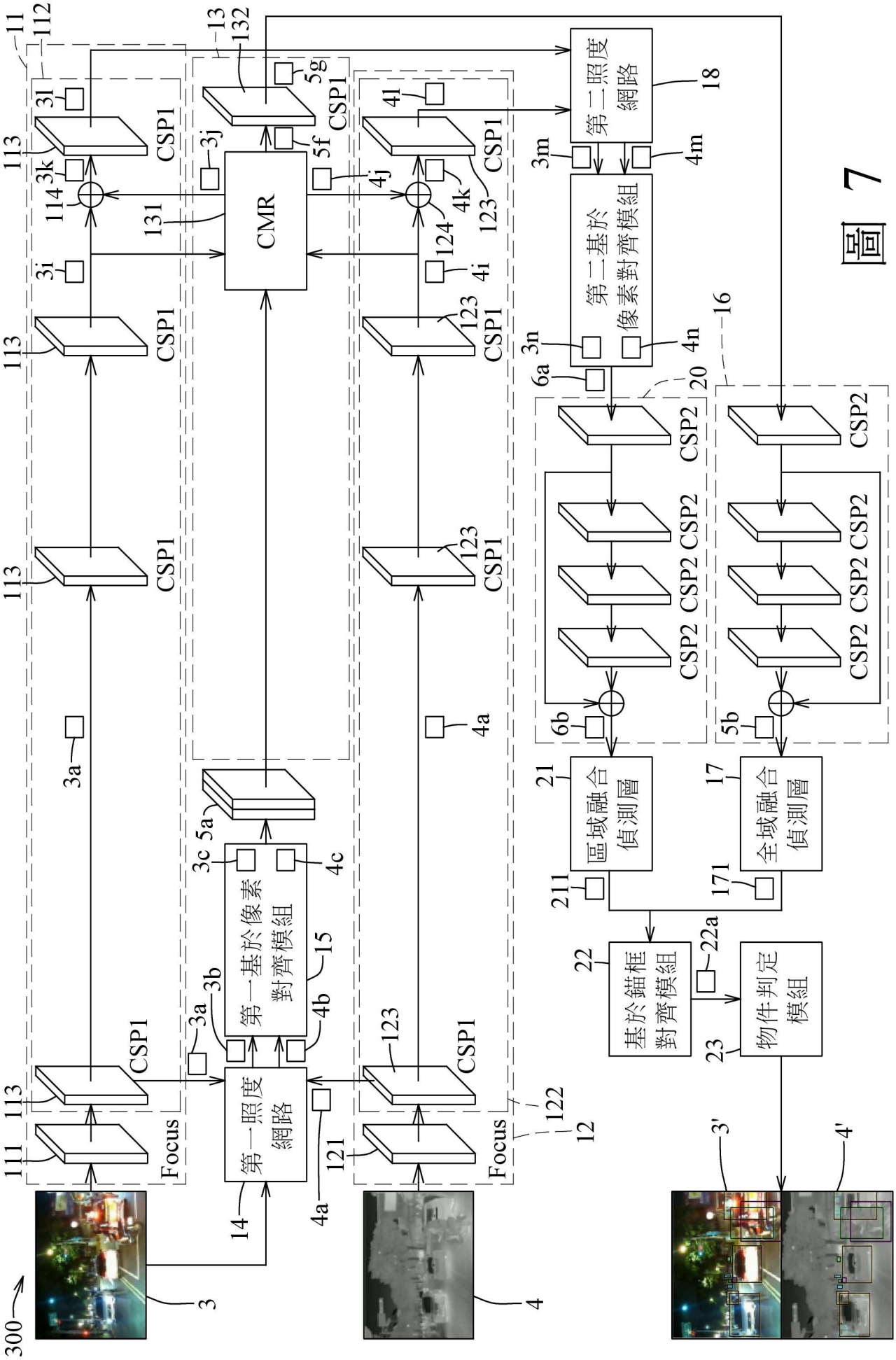
300
3
4
3'
4'
Focus
CSP1
CSP2
CMR
第一照度網路
第一基於像素對齊模組
第二照度網路
第二基於像素對齊模組
區域融合偵測層
全域融合偵測層
基於錨框對齊模組
物件判定模組

圖 7