Robust End-to-End Focal Liver Lesion Detection using Unregistered Multiphase Computed Tomography Images

Sang-gil Lee*, Eunji Kim*, Jae Seok Bae*, Jung Hoon Kim, and Sungroh Yoon, Senior Member, IEEE

*Equal contribution

Abstract—The computer-aided diagnosis of focal liver lesions (FLLs) can help improve workflow and enable correct diagnoses; FLL detection is the first step in such a computer-aided diagnosis. Despite the recent success of deep-learning-based approaches in detecting FLLs, current methods are not sufficiently robust for processing misaligned multiphase data. By introducing an attention-guided multiphase alignment in feature space, this study presents a fully automated, end-to-end learning framework for detecting FLLs from multiphase computed tomography (CT) images. Our method is robust to misaligned multiphase images owing to its complete learning-based approach, which reduces the sensitivity of the model’s performance to the quality of registration and enables a standalone deployment of the model in clinical practice. Evaluation on a large-scale dataset with 280 patients confirmed that our method outperformed previous state-of-the-art methods and significantly reduced the performance degradation for detecting FLLs using misaligned multiphase CT images. The robustness of the proposed method can enhance the clinical adoption of the deep-learning-based computer-aided detection system.

Index Terms—Computed tomography, Deep learning, Liver lesion, Object detection, Self-attention.

I. INTRODUCTION

HEPATOCELLULAR carcinoma (HCC) is the fifth most common cancer type and the second leading cause of cancer-related deaths worldwide [1]. Compared with other malignancies, HCC is unique in that it can be diagnosed based on imaging without histologic confirmation in patients with a high risk of developing HCC. Noninvasive diagnostic criteria established by major guidelines include hyperenhancement in the arterial phase and washout in the portal or delayed phase images [2-5]; these are considered to be associated with hemodynamic changes that occur during hepatocarcinogenesis [5]. Therefore, it is essential to obtain and interpret multiphase images of dynamic contrast-enhanced computed tomography (CT) or magnetic resonance images after the injection of contrast media to diagnose HCC.

When a radiologist diagnoses a focal liver lesion (FLL) in real-world clinical practice, he or she performs a comprehensive assessment of the lesion by characterizing its dynamic enhancement pattern on multiphase images. First, the precontrast, arterial, portal, and delayed phase images are aligned at a level where the lesion is best represented. Thereafter, the enhancements of the lesion, surrounding liver, and adjacent vascular structures on multiphase images are compared. If the lesion demonstrates a radiological hallmark of HCC (i.e., arterial enhancement and portal or delayed washout) and it is discovered in a patient with chronic liver disease, then a noninvasive diagnosis of HCC can be performed [6].

This complex process of detection and characterization using multiphase images differs from the traditional application of deep learning in detecting FLL using only single-phase images [2-4]. The evaluation of multiphase images is not merely the detection of a lesion, i.e., the alignment of images in multiple phases is required. If multiphase images are misaligned because of motion artifacts or respiratory movements, the diagnostic performance of deep learning would be undermined. In the detection task, wherein the model should accurately locate the lesion from images, the performance is significantly affected by misregistration because the model attempts to predict the bounding box based on an incorrectly aligned feature grid. Therefore, it is critical that the model accurately addresses the misalignments in multiphase images.

Existing literature regarding deep-learning-based systems presents challenges in processing multiphase images using external registration methods, and the deep-learning-based
model is assumed to operate on a perfectly aligned image space \cite{15,18}. Although this two-step approach renders the task easier to the deep-learning-based model, its implementation in real-world clinical practice presents several disadvantages: the diagnostic pipeline becomes complicated and the selection of registration method becomes a confounding factor for the model’s performance. Achieving multiphase registration using external tools often requires time-consuming optimization processes and domain-specific expertise, rendering the deployment of such diagnostic models to the real-world clinical environment challenging.

This study presents a fully-automated, end-to-end deep-learning-based computer-aided detection system that detects FLLs from multiphase CT images, leveraging a high-performance single-stage object detection model. The proposed model achieves a significantly higher performance compared to previous methods when detecting FLLs. More importantly, it provides robustness to misalignment, which has yet to be addressed in the literature. This is achieved by our novel trainable method, i.e., attention-guided multiphase alignment in a deep feature space. The model can leverage more extensive information present in multiphase CT images through a careful injection of parameterized task-specific inductive bias directly into neural network architectures. The method combines several neural subnetworks from distinct design disciplines to enable an inductive functionality required for the task. A grouped convolution \cite{19} decouples the feature extraction of multiphase images. A self-attention mechanism \cite{20} localizes and selects the salient aspect for detection. A deformable convolution kernel \cite{21,22} compensates for the spatial and morphological variances inherent to multiphase CT.

We provide an extensive and ablative analysis of the proposed method used in single-stage object detection models \cite{23,24}, using a custom-designed large-scale dataset containing multiphase CT images of 280 patients with HCC. Experimental results demonstrate that the method significantly outperforms previous state-of-the-art models for detecting FLLs under various conditions of registration quality.

II. RELATED STUDIES

A. Computer-aided Liver Lesion Diagnosis

Several deep-learning-based segmentation and detection approaches have been developed to detect FLLs based on portal phase CT images \cite{8,14}. \cite{14} presented a system developed using a state-of-the-art object detector (RetinaNet \cite{25}), introducing an anchor optimization method and utilizing dense segmentation masks approximated from weaker labels called Response Evaluation Criteria in Solid Tumors (RECIST). However, their system was evaluated on the single-phase DeepLesion \cite{26} dataset, which does not challenge the accurate handling of multiphase data.

Some studies proposed an end-to-end detection of FLLs with multiphase CT using grouped convolutional layers and recurrent neural networks for inter-phase feature extraction \cite{15-17}. In \cite{15}, a grouped single-shot multibox detector (GSSD) is proposed using an explicit decoupling feature extraction with a grouped convolution \cite{19}. This method has been successfully adapted to different neural architectures \cite{17}. However, it does not address the misalignment between modalities from the perspective of the model architecture. The existing approaches \cite{15-17} rely on an existing registration technique, that is, a preprocessing \cite{27}, and assume a perfect intermodal alignment. Hence, they do not provide robustness to misalignments, and the performance of the model cannot be guaranteed for data with an unknown registration quality. To our best knowledge, this study is the first to present an end-to-end, deep-learning-based detection of FLLs from a large-scale multiphase, unregistered CT database.

B. Attention Mechanism

The concept of attention has recently become a cornerstone in deep learning research \cite{28}, and the mechanism has been integrated into a wide range of tasks \cite{29-31}. In terms of medical images, the attention mechanism provides a trainable module that dynamically selects salient features present in an image that are effective for specified tasks \cite{32,33}. A soft-attention gating module proposed in \cite{32} constructs an attentive region in the current feature using the feature from a different scale, which is integrated to Sononet \cite{36} for image classification and U-net \cite{37} for image segmentation \cite{38}.

The previous studies in medical images have used the attention mechanism to ensure that the network focuses on the target area (e.g., lesion, viscera) in a single-phase image. In contrast, this study utilizes the attention mechanism to exploit the salient features from multiphase images and induce the model to guide the learning-based interphase alignment and cross-modal feature aggregation for the detection task.

C. Multimodal Registration

Medical image registration is crucial to achieving an accurate and robust computer-aided diagnosis system. Various approaches have been proposed, including statistical \cite{39,41} and learning-based ones \cite{42,43}, which aim to find an optimal transformation between a pair of images. By contrast, the multiphase registration task that we posed to the model is implicit in the deep feature space, not in the image space. This study does not directly provide the model with either the reference phase information or the registration-focused loss function. The implicit task is inherent in the dataset, wherein the model is required to dynamically focus on the misaligned information and adjust the offsets of the deep feature map to achieve the maximum object detection performance.

III. MULTIPHASE DATA

A. 3D CT Data Collection

We searched the image and pathology database of Seoul National University Hospital and retrospectively gathered data of patients who had undergone liver protocol CT before surgery for suspected HCC and were pathologically confirmed to have HCC based on surgical specimens. A board-certified radiologist reviewed the CT images of all the patients and included patients who had single HCCs with the typical enhancement pattern of HCC (i.e., arterial phase enhancement and portal/delayed
phase washout). Finally, 280 patients (233 men; mean age ± standard deviation, 58.3 ± 10.4 years) with single HCCs were determined (mean size ± standard deviation, 4.3 ± 2.7 cm; range 1-16 cm). All the CT scans were performed using multidetector CT (MDCT) scanners available at Seoul National University Hospital (n = 213) or various MDCT scanners at another hospital (n = 67). The protocol for the liver CT examinations comprised four phases: precontrast, arterial, portal, and delayed phases. After obtaining the precontrast images, 120 mL of 370 mg/mL of iopromide (Ultravist 370; Schering Korea, Seoul, Korea) was injected at a rate of 3.0–4.0 mL/s using a power injector (Envision CT; Medrad, Pittsburgh, PA, USA). Using the bolus tracking method, the arterial phase images were acquired automatically 19 s after an attenuation of 100 Hounsfield units was achieved in the descending aorta. The portal venous and delayed phase images were obtained 33 and 180 s after the arterial phase, respectively, after a contrast agent was administered. The entire dataset was approved by the institutional review board of the Seoul National University Hospital. The requirement for informed consent was waived.

B. Axial Alignment and Image Preprocessing

A custom multiphase dataset is generated from axial slices of 3D 512×512 pixel CT scan with variable heights per patient. Similar to the CT image preprocessing pipeline in [9], the Hounsfield unit range of the raw DICOM images is reduced to [-150, 250] by clipping, which is suitable for detecting hepatic lesions [46]. After the windowing, the images were normalized to have values between [0, 1] to be used as inputs to the model.

The axial coordinates of all four phases are manually aligned to ensure that all the phases indicated the same axial location of the 3D CT tensor. Subsequently, three consecutive axial CT slices were stacked (i.e., 2.5D images typically used in deep-learning-based approaches) for each of the four phases as a single data point. This was performed such that the model can leverage the axial features in the original images (e.g., to differentiate between blood vessels and lesions). This resulted in each data point containing a 12-channel image and a total of 4,748 data points. Fig. [1] shows some examples of the preprocessed data.

C. Bounding Box Label Annotation

A board-certified radiologist manually drew segmentation masks for HCCs on multiphase CT images using a graph cut-based [47] commercially available software (MEDIP; MEDI- CAL IP; Seoul, Korea). Among the contrast-enhanced phases, the one in which the HCC demonstrated the highest lesion-to-liver contrast was selected for placing the segmentation masks according to the radiologist’s discretion. We opted to construct the following numbers of segmentation masks for the phases: delayed (n = 158), arterial (n = 91), and portal (n = 31).

Besides, a post-annotation filtering with Gaussian blur is applied to the segmentation mask because the graph cut-based segmentation resulted in a fuzzy contour. We observed that a Gaussian kernel measuring 11×11 pixels resulted in the highest contour quality. As this study had formulated the task as an object detection problem, the bounding boxes were extracted from the segmentation mask contours using OpenCV [48] Python package.

As mentioned above, discrepancies were observed in the liver morphology and HCC location among the multiphase images, mainly due to respiratory movement. Consequently, the segmented mask for HCC in a particular phase did not perfectly match the same lesion depicted on the other phases (Fig. [1]). Hence, our dataset posed additional challenges for the model in terms of the phase that the model should learn to select the best phase for the four-phase input to achieve maximal performance.

IV. Method

This study presents a fully automated end-to-end learning approach for FLL detection from multiphase CT images. The following sections first describe the task-specific functionalities and inductive biases required for achieving maximal model performance on our real-world CT dataset. Subsequently, we present a trainable module that satisfies the desired properties. Finally, we present methods to incorporate the proposed modules into various high-performance single-stage object detection models.

This study was approved by the institutional review board (IRB No. 1704-175-849) of our institution, with a waiver of the requirement for informed consent.

A. Incorporating Parameterized Task-specific Priors

Standard approaches for deep-learning-based object detection use unimodal images as data distribution, such as a single natural image in the RGB color space. On the contrary, our FLL
The proposed method highlights two main advancements. First, we introduce the dynamic interphase extraction of spatial saliency from multistream features for lesion detection based on the self-attention mechanism \cite{20, 49}. Second, we propose a novel mechanism of attention-guided multiphase spatial alignment based on a deformable convolution kernel \cite{21, 22}. The following sections provide implementation details of the proposed method and the integration of the modules to high-performance object detection models.

B. Interphase Saliency Extraction with Self-attention

This section introduces the self-attention mechanism for extracting interphase spatial saliency. We adopted a modified version of convolutional self-attention layers \cite{49} that operated on the multiphase feature space (Left in Fig. 2). The layer was applied at two locations: one in the grouped feature map from the base CNN, and the other before the multiphase fusion with the 1 × 1 convolution (Fig. 3). Hence, the grouped feature map of the base CNN can communicate between previously decoupled feature channels inside the self-attention (SA) module. This is further leveraged by global-level spatial dependencies across the feature map from the non-local nature of self-attention. This enables the model to address the salient phase of the multiphase inputs from both the feature extraction and channel fusion stages, thereby enhancing the performance of lesion detection.

Assume that we have a deep image feature map \( x \in \mathbb{R}^{C \times N} \), where \( C \) is the number of feature channels, and \( N = H \times W \) is a flattened two-dimensional (2D) image feature. In our model, \( C \) was obtained from the concatenated features with \( C/4 \) channels, each via grouped convolutions. \( x \) was first transformed into three features, i.e., query \( q = W_q x \), key \( k = W_k x \), and value \( v = W_v x \), where \( W_q \in \mathbb{R}^{C \times C} \), \( W_k \in \mathbb{R}^{C \times C} \), and \( W_v \in \mathbb{R}^{C \times C} \) were parameterized by separate \( 1 \times 1 \) convolutions. \( C \) and \( \hat{C} \) are the number of bottleneck channels during the self-attention, where we used \( \hat{C} = C/8 \) and \( \hat{C} = C/2 \). The global-level self-attention map \( \beta \in \mathbb{R}^{N \times N} \) is obtained as follows:

\[
\beta_{ij} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}, \quad \text{where} \quad s = q^T k. \tag{1}
\]

The self-attention map \( \beta \) exhibits global dependencies between features, where \( \beta_{ij} \) corresponds to the extent to which a \( j^{th} \) feature location attends to an \( i^{th} \) location. From the self-attention map, a self-attention gate map \( g \in \mathbb{R}^{C \times N} \) is obtained as follows:

\[
g = v \beta^T. \tag{2}
\]

Finally, a residual self-attention feature map \( y \in \mathbb{R}^{C \times N} \) is obtained as follows:

\[
y = \sigma o + x, \quad \text{where} \quad o = W_o g \quad \text{and} \quad W_o \in \mathbb{R}^{C \times \hat{C}}. \tag{3}
\]
σ is a trainable parameter that serves as a gating mechanism for residual self-attention. It is initialized to zero such that the architecture can focus on the local dependencies first and then to learn to expand its focus to the non-local features. Consequently, the previously decoupled feature maps can communicate between phases by observing the global context (both channel- and location-wise) via the gated self-attention map. The same pooling was applied to value maps, causing the model to adaptively attend to the salient features. Additionally, the same self-attention mechanism based on the target architecture. This enables the model to attend to aggregated pixels instead of individual pixels in the feature space, thereby achieving lower run-time memory consumption and stabilized model dynamics. The dconv kernels have a trainable and dynamic receptive field based on inputs obtained by predicting the offsets to the original sampling grid by a separate convolutional layer. Subsequently, the deformable convolution filter operates on the augmented grid predicted by the offsets. Concretely, define the regular sampling grid \( \mathcal{R} \) of a 3 × 3 convolution filter with dilation 1 as \( \mathcal{R} = \{(−1, −1), (−1, 0), \ldots, (0, 1), (1, 1)\} \). The regular convolution filter on this fixed grid \( \mathcal{R} \) operates as follows:

\[
y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n),
\]

where \( p_0 \) is the output location from the convolution filter \( w \) applied to input \( x \).

The offset predictor computes \( \{Δp_{(n=1,...,N)}\} \), where \( N = |\mathcal{R}| \), through the separate convolution layer. Subsequently, the deformable convolution filter operates on the augmented grid as follows:

\[
y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n + Δp_n).
\]

The conventional application of the dconv layer predicts the same offsets across all channels. In contrast, because our desired functionality was the interphase feature-space alignment, the deformable convolution (DC) module is tasked to predict the four separate offsets corresponding to each grouped feature map. Subsequently, we applied dconv kernels with multiple offsets targeted differently for each grouped feature channel.

**C. Attention-guided Multiphase Alignment**

We endowed the system with robustness to registration mismatch by introducing a novel attention-guided multiphase alignment module in the feature space. This was achieved by incorporating a modified version of the deformable convolution (dconv) \[21, 22\] kernels into the network. We propose applying dconv kernels to the deep grouped feature map extracted by the backbone CNNs, whose alignment operations are guided by the SA module.

Fig. 3. Schematic diagram of GSSD++ (top) and PixelLink++ (bottom). Solid lines drawn along the surface of tensors indicate a feature map constructed by grouped convolutions. SA module refers to self-attention module and DC module refers to deformable convolutions.
The DC module is inserted after the self-attended feature map of the first source layers (e.g., conv4_3) of the VGG16 architecture in the GSSD model. The main objective is to have the model adaptively align the mismatch of offsets and morphological variances between phases from a fine-grained grouped feature map earlier in the base CNN forward pass. The DC module should receive a grouped feature map with a deep representation that is sufficiently rich to fully leverage the dconv filters and avoid overfitting. We observed that inserting the AC and DC modules into conv4_3 yielded the best results, i.e., better loss and detection metrics, while simultaneously minimizing overfitting.

The dconv kernels with multiphase offset predictors require relevant visual cues to possess the functionality of the intermodal alignment. This is achieved by guiding the learning process of the dconv kernels by leveraging the modified self-attention module described in the preceding section. Contrary to [24], where the method forward-propagates only the self-attention feature map, we propose to further leverage the self-attention gate map as an additional visual cue for the attention-guided multiphase alignment. Specifically, the gated self-attention map is concatenated with the feature map and processed by the DC module (Fig. 3). By receiving both the features and self-attention gate map as the inputs, the dconv filters can adaptively predict the geometric offsets for the multiphase features, with the global-level self-attention map as an additional context. The self-attention gate map can additionally receive gradients from the dconv layers directly, providing a shortcut for cooperative learning between two modules. We observed that this novel combination significantly improved the detection performance on the misregistered multiphase dataset.

V. EXPERIMENTAL RESULTS

The proposed method can be easily applied to various single-stage detection networks. To demonstrate its applicability, this study applied the proposed modules to the GSSD [15] and PixelLink [24], which are referred to as GSSD++ and PixelLink++, respectively; subsequently, we performed a comparative and ablative analysis using recent state-of-the-art FLL detection methods [14, 15, 17].

A. Implementation Details

We used VGG16 [50] as their backbone CNN architecture to accurately assess the performance of each model by decoupling the effectiveness of the proposed modules from incorporating more powerful base CNN architectures [51]. The downsample factor $D$ for the pooling operation inside the self-attention mechanism in Section IV-B is applied as $D = 1$ for GSSD++ and $D = 2$ for PixelLink++ from the hyper-parameter sweep over $D = \{1, 2, 4, 8\}$. For evaluation, the custom-curated 280-patient dataset is randomly split into 216, 54, and 10 patients respectively; subsequently, we performed a comparative and ablative analysis using recent state-of-the-art FLL detection methods [14, 15, 17].

B. Evaluation Metrics

We quantitatively measured the performance of each model based on the average precision (AP) over either the intersection over union (IoU) or intersection over bounding box (IoBB), which is a standard benchmark for deep-learning-based object detection models on natural images. The low-confidence predictions below 0.2 are filtered out to focus on the lesion detection performance of high-confidence box predictions and compare the performance of various models fairly.

C. Baseline Results

We measured the effectiveness of using multiphase CT images by training the baseline model using only the portal
phase images (Table I). The model performed significantly worse when trained with only portal phase images than with the multiphase baseline. This verifies that the GSSD leverages multiphase data effectively, consequently verifying the requirement for robustness to the multiphase registration mismatch of the model. Because the original architecture from the improved RetinaNet [14] and PixelLink [24] do not consider multiphase data, we adopted a grouped convolutional layer and a $1 \times 1$ convolutional channel fusion layer introduced in [15] to those models and confirmed that the method enabled the model to leverage multiphase data more effectively.

RetinaNet [25] and its optimized variant for liver lesion detection [14] are based on Feature Pyramid Networks (FPNs) [53]. Although the model is often considered as a state-of-the-art model, we did not observe any significant performance gain compared with other baselines when training was performed using bounding box annotations. The model demonstrated a clear performance improvement only when a dense segmentation mask label was used additionally (instead of the RECIST-based mask label used by [14], which is not accessible in our data). We refer to the model trained with stronger supervision of the segmentation mask as a strong baseline in our experimental result.

PixelLink [24] cast the bounding box regression problem into instance segmentation and used a balanced cross-entropy loss for a small area, enabling PixelLink-based models to excel at detecting smaller-sized lesions from the validation set. However, owing to the noisy post-filtering approach for the bounding box extraction from the predicted segmentation map, the performance degraded for metrics with a higher overlap threshold. SSD [23] and RetinaNet [25], which used a more conventional approach involving the use of a predefined set of densely distributed anchor boxes [54], achieved better performances on metrics with higher thresholds.

D. Improvements from the proposed method

Table I summarizes the detection performance of the proposed framework, GSSD++ and PixelLink++. As shown, the proposed models significantly outperformed the baseline models for all the metrics with various overlap thresholds. GSSD++ achieved an AP of 77.22% on IoBB50, i.e., 9.89%p higher than the GSSD. PixelLink++ achieved an AP of 73.52% on IoBB50, i.e., 9.08%p higher than PixelLink with grouped convolution and a channel fusion layer. Fig. 4 shows a qualitative evaluation of the liver lesion detection performances of GSSD++ and PixelLink++ compared with the baselines. Our models detected lesions with higher confidence (marked in red) and indicated fewer false positives compared with the baseline models, which localized the wrong location as the lesion. For $2^{nd}$, $3^{rd}$, and $6^{th}$ columns in Fig. 4, the baseline models detected an area as a lesion larger than the actual lesion. As shown in Fig. 1, a spatial mismatch occurred between the phases, which made it difficult for models without an internal mismatch ability to find accurate lesion regions.

E. Robustness to registration mismatch

To verify the robustness of the proposed method to a registration mismatch, we trained all models using the registered dataset and measured the performance degradation owing to such a mismatch. The data is registered using three different methods: rigid, affine, and non-rigid, by using the SimpleElastix [55] open-source package with the default parameters.
TABLE II

| Methods       | Dice Score |
|---------------|------------|
| Unregistered  | 0.9001     |
| Rigid         | 0.9119     |
| Affine        | 0.9213     |
| Non-rigid     | 0.9622     |

Fig. 5. Assessment of robustness to registration mismatch. Each model was trained using the registered version of a multiphase dataset by applying various methods from SimpleElastix. Here, † denotes that segmentation masks were used during training. A lower sensitivity indicates higher robustness to a registration mismatch.

and considering the portal phase as a fixed image. Each 2D slice image and lesion segmentation mask was registered on the portal phase image, and the bounding box was then extracted from the registered segmentation mask contours. Table II shows the level of registration mismatch of each dataset, which was measured with an overlap of the liver obtained from a liver segmentation network. Motivated by [56], we calculated the Dice score [57] between the liver from two different phases. For reliability, the Dice scores were averaged except for those cases in which the values were zero.

We defined the sensitivity of each model configuration to a registration mismatch as the ratio of performance degradation in the unregistered dataset when compared with the performance on each registered dataset:

\[
\text{sensitivity} = 1 - \frac{\text{perf.(} f_{\text{unregistered}}, D_{\text{unregistered}})\text{}}{\text{perf.(} f_{R}, D_{R})},
\]

where \( D_{\text{unregistered}} \) is the unregistered dataset and \( D_{R} \) is the dataset registered with method \( R \). In addition, \( \text{perf.(} f_{R}, D_{R}) \) is the performance of model \( f \) on dataset \( D_{R} \), which is trained with dataset \( D_{R} \).

Fig. 5 shows the average sensitivity of each model configuration for various levels of mismatch based on all metrics reported in Table II (AP with IoU30, IoU50, IoU70, IoBB30, IoBB50, and IoBB70 on the validation set, and IoU50 and IoBB50 on the test set). Because the registered data alleviated the modeling difficulties for the task, all models exhibited a performance degradation (positive sensitivity). Because the dataset registered with a non-rigid method was the best aligned dataset (Table II), the sensitivities of all models on \( D_{\text{non-rigid}} \) were larger than those on \( D_{\text{affine}} \) and \( D_{\text{rigid}} \). A model can be regarded as more robust to a registration misalignment if it has low sensitivity owing to less degradation in its performance. We discovered that the sensitivities of GSSD++ and PixelLink++ were significantly lower than those of the baseline models. The results verified that the proposed method is more generalizable than previous approaches owing to its robustness to a registration misalignment between phases.

The detection performance of each model is reported in the supplementary material.

F. Qualitative Analysis of Model Dynamics

We qualitatively examined the model behavior of GSSD++ by visualizing a self-attention map and the augmented receptive fields of the dconv kernels. Fig. 6 shows the self-attention gate map overlaid on the inputs. As shown, the SA module functioned according to its location based on the different inductive biases encouraged by the model. In the base architecture, the
gate map augmented by the self-attention weight was primarily focused on the shape of the organ, thereby providing enhanced visual cues for multiphase alignment and extracting saliency from the organ. In the fusion layer, self-attention successfully localized the salient aspects of the lesion by focusing on the characteristic textures present in the HCCs. The self-attention gate map is computed in a channel-wise manner from the global-level self-attention map (β in Equation 1). This can be augmented based on the context present on each feature channel as well as guide the learning of an intermodal alignment and saliency extraction of lesions in an adaptive manner.

Fig. 7 shows the dynamically augmented, phase-wise sampling grid of the dconv kernel. Here, the model adjusted the offsets for the precontrast and portal phase kernels in the left and right examples for the heavily deformed area of the liver, respectively, to align the multiphase deep features across phases. By adaptively correcting the registration mismatches inherent to the dataset guided by self-attention, the model can extract aligned deep feature maps more accurately and thereby achieve a robust object detection.

### G. Ablation Analysis

We performed an ablation analysis of GSSD++ and PixelLink++ to clearly demonstrate the superior performance yielded by the proposed modules. Compared with the final model configuration, modified variants obtained by removing certain modules from the final model indicated degraded performances with various degrees (Table III).

The dynamic phase-wise offset predictor, which allowed the features of each phase to be aligned, aided increase the detection performance in comparison with that afforded by a global offset predictor. Completely removing the DC module degraded the performance, which deteriorated even further because the model became more inclined to operate on a regular convolution sampling grid rather than a dynamic one. This shows the effectiveness of the alignment in the feature space through the DC module. To create the model variants without an interphase saliency extraction, we applied grouped convolutions [19] in the SA module (“w/o Interphase Attention” in Table III). Removing the interphase saliency extraction resulted in no information exchange between phases, which degraded the performance. This shows the effectiveness of saliency extraction through interphase communication. When the model lost its self-attention capability by the removal of corresponding modules (on both the base CNN and channel fusion locations), the performance was negatively affected. Even when a dynamic sampling grid was present, the model had to detect lesions from the equally weighted feature map without attending to the salient aspects, thereby resulting in a significant performance loss. The ablation result further highlights the importance of the novel method for attention-guided intermodal alignment to achieve the best lesion detection performance.

### VI. Conclusion

This study presented a fully automated, end-to-end liver lesion detection model that leveraged high-performance deep-learning-based single-stage object detection models. The proposed method was achieved by directly incorporating task-specific functionalities and inductive biases as parameterized modules into the model architecture, utilizing interphase self-attention and phase-wise deformable convolution kernels. We demonstrated a significantly improved detection performance by applying the proposed methods to two fundamentally different approaches for object detection, namely the GSSD [15] and PixelLink [24]; this implies that the proposed method is transferable to various single-stage detection models. Extensive experiments on both unregistered and registered real-world CT data revealed that the model outperformed previous state-of-the-art lesion detection models through inductive biases targeted at challenges inherent to real-world multiphase data.

The method is applicable to real-world multiphase CT data because the model does not require a perfect registration, which complicates the pipeline. The reduced sensitiveness and reliance on external registration methods would render the deployment of the model into real-world environments significantly more manageable. We believe that simplifying the process towards an end-to-end and scalable approach would enhance the clinical adoption of the deep-learning-based computer-aided detection system.

### References

[1] S. Mittal and H. B. El-Serag, “Epidemiology of hcc: consider the population,” Journal of clinical gastroenterology, vol. 47, p. S2, 2013.

[2] European Association for the Study of the Liver and European Organisation for Research and Treatment of Cancer, “Easl–eortc clinical practice guidelines: Management of hepatocellular carcinoma,” Journal of Hepatology, vol. 56, no. 4, pp. 908–943, 2012.

[3] M. Omata et al., “Asia–pacific clinical practice guidelines on the management of hepatocellular carcinoma: a 2017 update,” Hepatology international, vol. 11, no. 4, pp. 317–370, 2017.
[4] J. K. Heimbach et al., “Aasld guidelines for the treatment of hepatocellular carcinoma,” *Hepatology*, vol. 67, no. 1, pp. 358–380, 2018.

[5] A. Kitao et al., “Hepatocarcinogenesis: multistep changes of drainage vessels at ct during arterial portography and hepatic arteriography—radiologic-pathologic correlation,” *Radiology*, vol. 252, no. 2, pp. 605–614, 2009.

[6] J.-Y. Choi et al., “Ct and mr imaging diagnosis and staging of hepatocellular carcinoma: part ii. extracellular agents, hepatobiliary agents, and ancillary imaging features,” *Radiology*, vol. 273, no. 1, pp. 30–50, 2014.

[7] P. Bilic et al., “The liver tumor segmentation benchmark (lits),” *arXiv preprint arXiv:1901.04056*, 2019.

[8] W. Li et al., “Automatic segmentation of liver tumor in ct images with deep convolutional neural networks,” *Journal of Computer and Communications*, vol. 3, no. 11, p. 146, 2015.

[9] A. Ben-Cohen et al., “Fully convolutional network for liver segmentation and lesions detection,” in *Deep learning and data labeling for medical applications*. Springer, 2016, pp. 77–85.

[10] P. F. Christ et al., “Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 415–423.

[11] P. F. Christ et al., “Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks,” *arXiv preprint arXiv:1702.05970*, 2017.

[12] X. Han, “Automatic liver lesion segmentation using a deep convolutional neural network method,” *arXiv preprint arXiv:1704.07239*, 2017.

[13] X. Li et al., “H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes,” *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[14] M. Zlocha, Q. Dou, and B. Glocker, “Improving retinanet for ct lesion detection with dense masks from weak recist labels,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 402–410.

[15] S.-g. Lee et al., “Liver lesion detection from weakly-labeled multi-phase ct volumes with a grouped single shot multibox detector,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 693–701.

[16] D. Liang et al., “Combining convolutional and recurrent neural networks for classification of focal liver lesions in multi-phase ct images,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 666–675.

[17] D. Liang et al., “Multi-stream scale-insensitive convolutional and recurrent neural networks for liver tumor detection in dynamic ct images,” in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 794–798.

[18] R. Hasegawa et al., “Automatic segmentation of liver tumor in multiphase ct images by mask r-cnn,” in *2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech)*. IEEE, 2020, pp. 231–234.

[19] A. Krizhevsky et al., “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] A. Vaswani et al., “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[21] J. Dai et al., “Deformable convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[22] X. Zhu et al., “Deformable convnets v2: More deformable, better results,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9308–9316.

[23] W. Liu et al., “Ssd: Single shot multibox detector,” in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[24] D. Deng et al., “Pixellink: Detecting scene text via instance segmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[25] T.-Y. Lin et al., “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[26] K. Yan et al., “Deepleision: automated mining of large-scale lesion annotations and universal lesion detection with deep learning,” *Journal of Medical Imaging*, vol. 5, no. 3, p. 036501, 2018.

[27] C. Dong et al., “Non-rigid image registration with anatomical structure constraint for assessing locoregional therapy of hepatocellular carcinoma,” *Computerized Medical Imaging and Graphics*, vol. 45, pp. 75–83, 2015.

[28] D. Hassabnis et al., “Neuroscience-inspired artificial intelligence,” *Neuron*, vol. 95, no. 2, pp. 245–258, 2017.

[29] I. Sutskever et al., “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[30] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=B1xsqj09Fm

[31] J. Shen et al., “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[32] J. Schlemper et al., “Attention gated networks: Learning to leverage salient regions in medical images,” *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[33] J. Zhang et al., “Attention residual learning for skin lesion classification,” *IEEE transactions on medical imaging*, 2019.

[34] A. Sinha and J. Dolz, “Multi-scale self-guided attention for medical image segmentation,” *IEEE Journal of Biomedical and Health Informatics*, 2020.

[35] Y. Wang et al., “Abdominal multi-organ segmentation with
organ-attention networks and statistical fusion,” Medical image analysis, vol. 55, pp. 88–102, 2019.

[36] C. F. Baumgartner et al., “Real-time detection and localisation of fetal standard scan planes in 2d freehand ultrasound,” CoRR abs/1612.05601, 2016.

[37] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[38] O. Oktay et al., “Attention u-net: Learning where to look for the pancreas,” arXiv preprint arXiv:1804.03999, 2018.

[39] J. Ashburner, “A fast diffeomorphic image registration algorithm,” Neuroimage, vol. 38, no. 1, pp. 95–113, 2007.

[40] B. Glocker et al., “Dense image registration through mrfs and efficient linear programming,” Medical image analysis, vol. 12, no. 6, pp. 731–741, 2008.

[41] A. V. Dalca et al., “Patch-based discrete registration of clinical brain images,” in International Workshop on Patch-based Techniques in Medical Imaging. Springer, 2016, pp. 60–67.

[42] B. T. Yeo et al., “Learning task-optimal registration cost functions for localizing cytoarchitecture and function in the cerebral cortex,” IEEE transactions on medical imaging, vol. 29, no. 7, pp. 1424–1441, 2010.

[43] G. Balakrishnan et al., “Voxelmorph: a learning framework for deformable medical image registration,” IEEE transactions on medical imaging, vol. 38, no. 8, pp. 1788–1800, 2019.

[44] B. Kim et al., “Unsupervised deformable image registration using cycle-consistent cnn,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 166–174.

[45] M. C. Lee et al., “Image-and-spatial transformer networks for structure-guided image registration,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 337–345.

[46] W. W. Mayo-Smith et al., “Detecting hepatic lesions: the added utility of ct liver window settings,” Radiology, vol. 210, no. 3, pp. 601–604, 1999.

[47] Y. Boykov and G. Funka-Lea, “Graph cuts and efficient nd image segmentation,” International journal of computer vision, vol. 70, no. 2, pp. 109–131, 2006.

[48] G. Bradski, “The opencv library,” Dr. Dobb’s Journal: Software Tools for the Professional Programmer, vol. 25, no. 11, pp. 120–123, 2000.

[49] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[50] K. He et al., “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[51] L. R. Dice, “Measures of the amount of ecologic association between species,” Ecology, vol. 26, no. 3, pp. 297–302, 1945.
Sang-gil Lee received the dual B.S. degree in electrical and computer engineering, and applied biology and chemistry from Seoul National University, Seoul, South Korea, in 2016, where he is currently pursuing the integrated M.S./Ph.D. degree in electrical and computer engineering. His research interests include artificial intelligence, deep learning, and medical image computing.

Eunji Kim received the B.S. degree in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2018, where she is currently pursuing the integrated M.S./Ph.D. degree in electrical and computer engineering. Her research interests include artificial intelligence, deep learning, and computer vision.

Jae Seok Bae received the B.S. degree in biological sciences from Korea Advanced Institute of Science and Technology, South Korea, in 2008. He received the M.D. degree from the College of Medicine, Seoul National University, South Korea, in 2012, and M.S. and Ph.D. degrees from Seoul National University, in 2017 and 2020, respectively. He is currently a Clinical Assistant Professor with the Department of Radiology, Seoul National University Hospital.

Jung Hoon Kim received the M.D. degree from the College of Medicine, Hanyang University, South Korea, in 1994, and M.S./Ph.D. degree from Catholic University, South Korea, in 2005. He was an Assistant Professor with the Department of Radiology, Soonchunhyang University Hospital, South Korea, from 2005 to 2008, and a Visiting Professor and Visiting Assistant Professor with the Department of Radiology and Body Imaging, University of Iowa Hospitals & Clinics, Iowa City, from 2008 to 2009. He is currently a Professor with the Department of Radiology and Institute of Radiation Medicine, Seoul National University College of Medicine.

Sungroh Yoon (S’99–M’06–SM’11) received the B.S. degree in electrical engineering from Seoul National University, South Korea, in 1996, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, CA, USA, in 2002 and 2006, respectively, where he was a Visiting Scholar with the Department of Neurology and Neurological Sciences, from 2016 to 2017. He held research positions at Stanford University and Synopsys, Inc., Mountain View. From 2006 to 2007, he was with Intel Corporation, Santa Clara. He was an Assistant Professor with the School of Electrical Engineering, Korea University, from 2007 to 2012. He is currently a Professor with the Department of Electrical and Computer Engineering, Seoul National University, South Korea. His current research interests include machine learning and artificial intelligence. He was a recipient of the SNU Education Award, in 2018, the IBM Faculty Award, in 2018, the Korean Government Researcher of the Month Award, 2018, the BRIC Best Research of the Year, in 2018, the IMIA Best Paper Award, in 2017, the Microsoft Collaborative Research Grant, in 2017 and 2020, the SBS Foundation Award, in 2016, the IEEE Young IT Engineer Award, in 2013, and many other prestigious awards. Since February 2020, Prof. Yoon has been serving as the chairperson (minister-level position) of the Presidential Committee on the Fourth Industrial Revolution established by the Korean Government.
Appendix

A. DETAILS IN 3D CT DATA COLLECTION

All the CT scans were performed using multidetector CT scanners available at Seoul National University Hospital (n= 213): 320-MDCT (Aquilion ONE, Canon Medical Systems, Tokyo, Japan) (n = 9), 64-MDCT (Brilliance 64, Ingenuity, or IQon, Philips Healthcare, Cleveland, OH, USA; Somatom Definition, Siemens Healthineers, Forchheim, Germany; or Discovery CT750 HD, GE Healthcare, Milwaukee, WI, USA) (n = 131), 16-MDCT (Sensation 16, Siemens Healthineers, Forchheim, Germany; or Mx 8000 IDT 16, Philips Healthcare, Cleveland, OH, USA) (n = 68), 8-MDCT (LightSpeed Ultra, GE Healthcare, Milwaukee, WI, USA) (n = 5), or various MDCT scanners at another hospital (n = 67). Because our data were retrospectively collected from a tertiary referral hospital, the images were obtained from the CT machines of various manufacturers. However, all CT examinations were performed using the routine protocol for liver CT, including those performed at other hospitals.

B. DETAILS IN TRAINING

We trained all models with a batch size of 32 until convergence was attained. We employed a stochastic gradient descent optimizer with a momentum of 0.9 and weight decay of $5 \times 10^{-4}$ to train the SSD- and PixelLink-based models. The models were trained for 60,000 training iterations, with a learning rate of $10^{-3}$. We reduced the learning rate by 0.1 times after 30,000 and 50,000 iterations. Furthermore, we reduced the learning rate for the dconv kernel by 0.1 to ensure numerical stability throughout the training process [22]. The improved RetinaNet [14] and MSCR [17] were trained based on the protocol of the corresponding open-source implementations.

We implemented the model using the PyTorch library and trained the models on an NVIDIA V100 GPU with 32 GB of VRAM.

C. DATA SPLIT

To perform a fair and unbiased model evaluation without data leakage, we split our custom dataset prior to the experiment without any manual data examination. Consequently, the values of the test set metrics were higher than those of the validation set metrics across all the experiments (IoU50 and IoBB50 in Tables I, II, S1, and S2). This was due to the 10 patients in the test set who had HCC sizes that were larger than the average size, which facilitated detection by all the model configurations. However, this is not a concerning factor as the test set metrics were consistent for all the models, and the relative scale was maintained throughout the configurations, indicating that none of the models had overfitted to the training set.

D. PERFORMANCE ON REGISTERED DATA

As described in Section V-E in main manuscript, we assessed the model’s robustness with a registered variant of our dataset using SimpleElastix [55] Python package. Fig. S1 shows some examples of the unregistered and registered data. Table S1 shows the detection performance of the models trained with the registered dataset. GSSD++ and PixelLink++ still retained the state-of-the-art performance on the registered dataset. Table S2 shows the sensitivity on each registered version of the dataset. The sensitivities of GSSD++ and PixelLink++ were significantly lower than those of the baseline models.

E. FPN-BASED DETECTORS

Although the proposed method can be incorporated into other architectures, the architecture based on the FPN [53] such as improved RetinaNet [14] must be modified significantly, thereby defeating the purpose of comparative evaluation. Specifically, our method performed the best with the deep feature representation at a resolution lower than that of the image space, whereas the FPN featured multiscale object prediction via a lateral connection from a shallow feature map with a significantly higher resolution. In fact, we observed a marginal gain in detection performance when we applied the method to FPN-based detectors. Integrating the proposed method via alternative modules into such an architecture would be an interesting future study.

3We used other training configurations as well, including ours, on these models; however, worse performances were obtained.
Fig. S1. Visualization of registration of multiphase CT dataset. CT images and segmentation masks are registered using the non-rigid method from SimpleElastix with default parameters. Portal phase image is considered as a fixed image. Each row represents a single data point. Green contour corresponds to segmentation masks and red box corresponds to bounding box annotations obtained from the segmentation mask for the object detection task. Left: examples of unregistered images. Right: examples of registered images. After registration, the ground truth labels become consistent with all phases.

TABLE S1
Detection performance of models trained using the registered version of a multiphase dataset. The performance of each model was evaluated based on the dataset applied to train the model. The dataset was registered using rigid, affine, and non-rigid methods from SimpleElastix. Here, † denotes that segmentation masks were used during training.

| Methods | Validation Set(%) | Test Set(%) |
|---------|-------------------|-------------|
|         | IoU30  | IoU50  | IoU70  | IoUU30 | IoU50  | IoU70  | IoU50  | IoU70  |
| Rigid Method |               |           |         |         |           |         |         |           |         |
| Improved RetinaNet [14] + G.Conv & C.Fusion | 60.57  | 51.11  | 33.52  | 63.94  | 61.70  | 55.00  | 62.30  | 70.40  |
| Improved RetinaNet [14] + G.Conv & C.Fusion† | 62.63  | 53.53  | 37.77  | 65.24  | 62.99  | 56.27  | 67.56  | 74.10  |
| GSSD [15] | 58.69  | 51.08  | 37.15  | 60.11  | 58.72  | 53.44  | 64.32  | 69.08  |
| MSCR [17] | 67.14  | 56.12  | 33.61  | 69.46  | 60.23  | 41.50  | 62.02  | 65.58  |
| PixelLink + G.Conv & C.Fusion | 73.60  | 56.39  | 18.92  | 77.28  | 64.77  | 30.13  | 59.07  | 65.12  |
| GSSD++ (Ours) | 62.37  | 53.62  | 40.70  | 67.71  | 65.46  | 60.37  | 68.14  | 77.99  |
| PixelLink++ (Ours) | 74.85  | 57.13  | 23.78  | 78.27  | 64.41  | 34.31  | 64.58  | 75.19  |
| Affine Method |               |           |         |         |           |         |         |           |         |
| Improved RetinaNet [14] + G.Conv & C.Fusion | 62.28  | 53.89  | 33.69  | 65.00  | 62.75  | 55.41  | 61.37  | 67.44  |
| Improved RetinaNet [14] + G.Conv & C.Fusion† | 64.30  | 54.11  | 35.26  | 67.42  | 65.78  | 57.29  | 66.81  | 75.40  |
| GSSD [15] | 59.17  | 51.17  | 35.44  | 60.84  | 58.84  | 53.74  | 64.87  | 69.33  |
| MSCR [17] | 70.12  | 53.47  | 20.38  | 74.50  | 60.81  | 30.20  | 60.43  | 64.47  |
| PixelLink + G.Conv & C.Fusion | 70.12  | 51.93  | 17.96  | 72.31  | 58.47  | 27.47  | 60.00  | 65.38  |
| GSSD++ (Ours) | 63.60  | 55.15  | 40.45  | 66.87  | 64.76  | 59.98  | 68.66  | 76.63  |
| PixelLink++ (Ours) | 73.10  | 59.83  | 25.20  | 76.21  | 66.87  | 35.33  | 65.43  | 72.30  |
| Non-rigid Method |               |           |         |         |           |         |         |           |         |
| Improved RetinaNet [14] + G.Conv & C.Fusion | 61.29  | 55.09  | 37.91  | 63.92  | 62.10  | 57.87  | 67.25  | 70.90  |
| Improved RetinaNet [14] + G.Conv & C.Fusion† | 64.66  | 56.71  | 38.24  | 66.82  | 66.78  | 61.29  | 69.89  | 76.75  |
| GSSD [15] | 63.68  | 55.19  | 38.08  | 66.84  | 65.54  | 60.24  | 67.46  | 75.97  |
| MSCR [17] | 68.82  | 59.75  | 33.38  | 70.96  | 63.39  | 41.67  | 61.56  | 66.60  |
| PixelLink + G.Conv & C.Fusion | 72.41  | 55.22  | 23.12  | 75.21  | 63.11  | 33.65  | 59.24  | 66.28  |
| GSSD++ (Ours) | 64.05  | 55.63  | 41.93  | 69.11  | 67.33  | 62.76  | 69.94  | 79.34  |
| PixelLink++ (Ours) | 75.74  | 58.46  | 24.73  | 78.89  | 66.13  | 35.35  | 65.64  | 75.10  |
TABLE S2
ASSESSMENT OF ROBUSTNESS TO A REGISTRATION MISMATCH. THE SENSITIVITY IS REPORTED. HERE, † DENOTES THAT SEGMENTATION MASKS WERE USED DURING TRAINING. A LOWER PERFORMANCE DEGRADATION INDICATES A HIGHER ROBUSTNESS TO A REGISTRATION MISMATCH.

| Methods                        | Validation Set(%) | Test Set(%) | Average |
|--------------------------------|-------------------|-------------|---------|
|                                | IoU30  | IoU50  | IoU70  | IoU50  | IoUBB30 | IoUBB50 | IoUBB70 | IoU50  | IoUBB50 |
| Rigid Method                   |        |        |        |        |         |         |         |        |         |
| Improved RetinaNet [14] + G.Conv & C.Fusion | 4.71   | 5.04   | -5.03  | 6.03   | 5.18    | 4.58    | -0.27   | 5.82   | 3.26    |
| Improved RetinaNet [14] + G.Conv & C.Fusion† | -1.32  | 0.78   | 5.57   | 0.04   | 0.23    | -0.66   | 6.07    | 8.06   | 2.35    |
| GSDD [15]                      | 3.59   | 14.41  | 2.00   | 3.12   | 2.62    | 0.69    | 6.63    | 2.54   | 4.45    |
| MSCR [17]                      | 5.09   | 8.53   | 16.26  | 3.39   | 4.53    | -0.06   | 10.41   | 8.49   | 7.08    |
| PixelLink + G.Conv & C.Fusion  | 5.97   | 9.82   | -5.18  | 6.95   | 13.42   | 1.63    | 2.00    | 1.04   | 4.46    |
| GSDD++ (Ours)                  | 1.36   | 1.34   | 1.52   | -0.27  | 0.40    | 0.63    | 0.40    | 0.99   | 0.80    |
| PixelLink++ (Ours)             | 0.98   | -1.12  | -0.96  | 0.78   | -1.38   | 0.92    | 0.60    | 2.22   | 0.26    |
| Affine Method                  |        |        |        |        |         |         |         |        |         |
| Improved RetinaNet [14] + G.Conv & C.Fusion | 7.33   | 9.95   | -4.53  | 7.57   | 6.75    | 5.29    | -1.79   | 1.70   | 4.03    |
| Improved RetinaNet [14] + G.Conv & C.Fusion† | 1.31   | 1.85   | -1.14  | 3.28   | 4.46    | 1.13    | 5.02    | 9.65   | 3.19    |
| GSDD [15]                      | 4.38   | 14.56  | -2.74  | 4.29   | 2.82    | 1.25    | 7.41    | 2.88   | 4.36    |
| MSCR [17]                      | 7.15   | 9.88   | 17.12  | 4.10   | 5.57    | 0.60    | 10.11   | 11.25  | 8.22    |
| PixelLink + G.Conv & C.Fusion  | 1.32   | 4.91   | 2.36   | 3.47   | 7.78    | 1.86    | 4.20    | 0.05   | 3.24    |
| GSDD++ (Ours)                  | 1.87   | 4.08   | 0.90   | -1.52  | -0.69   | -0.02   | 1.15    | -0.77  | 0.63    |
| PixelLink++ (Ours)             | -1.40  | 3.44   | 4.72   | -1.90  | 2.35    | 3.79    | 1.90    | -1.69  | 1.40    |
| Non-Rigid Method               |        |        |        |        |         |         |         |        |         |
| Improved RetinaNet [14] + G.Conv & C.Fusion | 5.82   | 11.91  | 7.12   | 6.01   | 5.78    | 9.31    | 7.11    | 6.49   | 7.44    |
| Improved RetinaNet [14] + G.Conv & C.Fusion† | 1.86   | 6.35   | 6.73   | 2.41   | 5.89    | 7.58    | 9.20    | 11.24  | 6.41    |
| GSDD [15]                      | 11.15  | 20.78  | 4.39   | 12.88  | 12.76   | 11.90   | 10.97   | 11.37  | 12.02   |
| MSCR [17]                      | 7.41   | 14.08  | 15.68  | 5.44   | 9.29    | 0.36    | 9.73    | 9.89   | 8.98    |
| PixelLink + G.Conv             | 4.43   | 7.91   | 13.93  | 4.39   | 11.14   | 11.92   | 2.28    | 2.78   | 7.35    |
| GSDD++ (Ours)                  | 2.56   | 4.91   | 4.41   | 1.77   | 3.16    | 4.41    | 2.96    | 2.67   | 3.36    |
| PixelLink++ (Ours)             | 2.14   | 1.18   | 2.91   | 1.56   | 1.26    | 3.85    | 2.21    | 2.10   | 2.15    |