A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images

PABLO MESSINA, PABLO PINO, DENIS PARRA, and ALVARO SOTO, Computer Science Department, Pontificia Universidad Católica de Chile, Chile
CECILIA BESA, SERGIO URIBE, and MARCELO ANDÍA, Department of Radiology, School of Medicine, Pontificia Universidad Católica de Chile, Chile
CRISTIAN TEJOS, Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Chile
CLAUDIA PRIETO, School of Biomedical Engineering and Imaging Sciences, King's College London, St Thomas' Hospital, UK
DANIEL CAPURRO, School of Computing and Information Systems, The University of Melbourne, Australia

Every year physicians face an increasing demand of image-based diagnosis from patients, a problem that can be addressed with recent artificial intelligence methods. In this context, we survey works in the area of automatic report generation from medical images, with emphasis on methods using deep neural networks, with respect to (1) Datasets, (2) Architecture Design, (3) Explainability, and (4) Evaluation Metrics. Our survey identifies interesting developments but also remaining challenges. Among them, the current evaluation of generated reports is especially weak, since it mostly relies on traditional Natural Language Processing (NLP) metrics, which do not accurately capture medical correctness.

CCS Concepts: • **Computing methodologies** → **Computer vision**; **Natural language generation**; Neural networks; • **Applied computing** → **Health care information systems;**

Additional Key Words and Phrases: Medical report generation, medical image captioning, natural language report, medical images, deep learning, explainable artificial intelligence

This work was funded by the National Agency for Research and Development (ANID) / Scholarship Program / Doctorado Becas Chile/2019 - 21191569 & Magíster Becas Chile/2020 - 22201476, Millennium Science Initiative Program, Code ICN17_002 (IMFD), ICN2021_004 (iHEALTH) and by Basal Fund for Center of Excellence FB210017 (CENIA). In addition, we thank Fondecyt grant 1191791.
Pablo Messina and Pablo Pino contributed equally to this research.
Authors' addresses: P. Messina, P. Pino, D. Parra, and A. Soto, Computer Science Department, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, 7820436, Santiago, Chile; emails: {pamessina, pdpino}@uc.cl, dparra@ing.puc.cl, asoto@uc.cl; C. Besa, S. Uribe, and M. Andía, Department of Radiology, School of Medicine, Pontificia Universidad Católica de Chile, Avda. Libertador Bernando O'Higgins 340, 8320000, Santiago, Chile; emails: besacecilia@gmail.com, {suribe, meandia}@uc.cl; C. Tejos, Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, 7820436, Santiago, Chile; email: ctejos@puc.cl; C. Prieto, School of Biomedical Engineering and Imaging Sciences, King's College London, St. Thomas' Hospital, Lambeth Palace Rd., SE1 7EH, London, UK; email: cdprieto@gmail.com; D. Capurro, School of Computing and Information Systems, The University of Melbourne, Level 8, Doug McDonell Building, 3010, Melbourne, Australia; email: dcapurro@unimelb.edu.au.

0360-0300/2022/09-ART203 $15.00
https://doi.org/10.1145/3522747

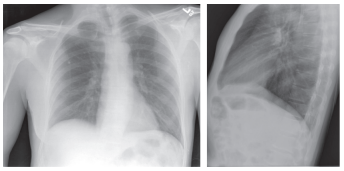
ACM Reference format:
Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images. *ACM Comput. Surv.* 54, 10s, Article 203 (September 2022), 40 pages.
https://doi.org/10.1145/3522747

1 INTRODUCTION

The rapid and successful development of deep learning in research fields such as Computer Vision [74] and **Natural Language Processing (NLP)** [101] has found an important application area in healthcare, sustaining the promise of a future with more efficient and affordable medical care. Research over the last 5 years shows a clear improvement in **computer-aided detection (CAD)**, specifically in disease prediction from medical images [37, 62, 110, 136, 142] as well as from **Electronic Health Records (EHRs)** [119], by using **deep neural networks (DNNs)** and treating the problem as a supervised classification or segmentation task. Recently, Topol [135] indicates that the need for diagnosis and reporting from image-based examinations far exceeds the current medical capacity of physicians in the United States. This situation promotes the development of automatic image-based diagnosis as well as automatic reporting. Furthermore, the lack of specialist physicians is even more critical in resource-limited countries [117], and therefore the expected impacts of this technology would become even more relevant.

However, the elaboration of high-quality medical reports from medical images, such as chest X-rays, **computed tomography (CT),** or **magnetic resonance imaging (MRI)** scans, is a task that requires a trained radiologist with years of experience. In this context, **deep learning (DL)** combined with other **Artificial Intelligence (AI)** techniques appears as a viable and promising solution to alleviate the physician scarcity problem, by both automating the report generation process and enhancing radiologists' performance through assisted report generation. AI is set to have a significant impact on the medical imaging market and, hence, how radiologists work, with the ultimate goal of better patient outcomes. The pace of research in this area is rapid, and to the best of our knowledge, previous surveys on this topic [6, 97, 104] do not cover aspects of explainability [46], medical correctness, and physician-centered evaluation. This article enhances these previous surveys by analyzing more than 20 additional works and datasets. Furthermore, unlike previous surveys, in this article we pay special attention to **explainable AI (XAI)**. XAI is a set of methods and technologies that will allow physicians to better understand the rationale behind automatic reports from black-box algorithms [45], potentially increasing trust for their actual clinical use.

Contribution. We summarize the state of research in automatic report generation from medical images. We perform an exhaustive review of the literature, consisting of 40 articles published in journals, conferences, and conference workshop proceedings. We first present an overview of the task (Section 2), followed by the survey methodology for search and selection of articles (Section 3), and the research questions driving this research (Section 4). We then analyze articles regarding four dimensions: Datasets Used (image modalities and clinical conditions, in Section 5.1), Model Design (standard practices, input and output, visual and language components, domain knowledge, auxiliary tasks, and optimization strategies, in Section 5.2), Explainability (Section 5.3), and Evaluation Metrics (Section 5.4). We also compare model performance of several articles (Section 5.5), identifying unsolved challenges across all reviewed articles and proposing potential avenues for future research (Section 6). Lastly, we discuss the limitations of this work (Section 7) and offer the main conclusions (Section 8). Our survey provides valuable insights to guide future research on automatic report generation from medical images.



Manual tags: Calcified Granuloma/lung/upper lobe/right
Automatic tags: Calcified granuloma

Comparison: Chest radiographs XXXX.
Indication: XXXX-year-old male, chest pain.
Findings: The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. Stable calcified granuloma within the right upper lung. No acute bone abnormality.
Impression: No acute cardiopulmonary process.

Fig. 1. Example from the IU X-ray dataset, frontal and lateral chest x-rays from a patient, alongside the natural language report and the annotated tags. XXXX is used for anonymization of the report.

2 TASK OVERVIEW

From a purely computational perspective, the following is the main task addressed by most articles analyzed in this survey: given as input one or more medical images of a patient, a text report is output that is as similar as possible to one generated by a radiologist. From a machine learning point of view, creating a system that performs such a task would require learning a *generative model* from instances of reports written by radiologists. Figure 1 presents one example of such a report, taken from the IU X-ray dataset [28]. We see two input X-ray images (frontal and lateral), below them some annotations (Tags)—some manually annotated by a radiologist and others automatically annotated—and on the right side the report with four different sections (comparison, indication, findings, and impression). If we consider the clinical workflow of generating a medical imaging report, several aspects should be taken into account before diving into a concrete implementation.

The first aspect is considering additional patient information in the process of report generation. Most of the time, the physician asking for medical imaging is the primary care physician or a medical specialist. This implies that when radiologists write a report, they generally have patient-relevant clinical information, usually provided in the section *Indication* as shown in Figure 1. Also, the *Comparison* section can provide information of a serial follow-up procedure, to evaluate the evolution of a patient over time (e.g., aneurysm, congenital heart disease). Then, one decision can be whether or not to use these *Indication* and *Comparison* data to generate the sections *Findings*, *Impression*, or both of them.

Second, the model for report generation should consider the diversity on medical images as well as body regions and conditions. There are several types of medical images, such as X-rays, CT, MRI, PET, and SPECT. This implies that a model for text report generation that deals with only one type of input medical image might not solve it for other types. Also, ideally, a model should be able to generate reports from different parts of the human anatomy and diverse medical conditions. To adequately achieve this task, different body regions must have a balanced and sizable training set. Many works surveyed in this article focus on one specific part of the body and particular illnesses, which limits the applicability of these methods to generalize to all possible diagnosis tasks.

Lastly, even if an AI system has perfect report generation accuracy, we might wonder if we can trust a machine in such a critical domain. One of the reasons for preferring a radiologist rather than an automated, highly accurate AI system is the chance of understanding the rationale behind the findings and impressions. In this sense, explainable AI [46] is of great importance in securing their adoption in a clinical setting.

Table 1. Articles Found for Each Query and Database and Included or Discarded with Different Criteria

Query	Google Scholar	PubMed	Scopus	ACM	WoK	IEEE Xplore	Springer	Total
1	32	1	19	2	9	7	13	34
2	21	2	20	2	11	3	18	37
Selected with inclusion criteria (all queries)								45
Discarded with exclusion criteria [5, 56, 93, 145, 147]								5
Total articles [7, 16, 36, 38, 40, 44, 47–49, 51, 61, 67, 68, 76, 85–87, 89, 92, 94, 95, 98, 120, 120, 123, 126, 128, 131, 132, 144, 148, 150, 151, 153–158, 158, 162, 163, 163]								**40**

WoK stands for Web of Knowledge. In both queries, only articles from journal, conference or conference workshops proceedings were included.
Query 1: (medical OR medicine OR health) AND "report generation" AND (images OR image).
Query 2: (medical OR medicine OR health) AND (images OR image) AND (report OR diagnostic OR description OR caption) AND (generation OR automatic) in ABSTRACT.
Relaxed queries: (medical report generation), (medical report image), (diagnostic captioning).

3 SURVEY METHODOLOGY: SEARCH AND SELECTION OF ARTICLES

To collect the articles reviewed, we performed three main steps: retrieval, selection, and exclusion. We further describe each step in the following paragraphs.

Study retrieval. To retrieve the articles, we used seven search engines, namely Google Scholar, PubMed, Scopus, ACM Digital Library, Web of Knowledge, IEEE Xplore, and Springer, and two specific queries, plus other more relaxed queries, described in Table 1. The relaxed queries returned articles already found with the two main queries. In this step we only considered journals, conferences, and conference workshop proceedings.

Study selection. Given the query results, a selection was performed applying inclusion criteria by reading the title, abstract, and keywords of each article. If there was uncertainty after reading these sections, we included the article for revision and decided afterward if it should be excluded with exclusion criteria. The inclusion criteria were the following: at least a part of the study focused on report generation from medical images. The images can be from any kind (e.g., X-ray, MRI scans, CT scans), must be from humans, and may include one or more pathologies of any type.[1] The report must be in natural language form, composed of least one or more sentences, and must be automatically or semi-automatically generated by a computational system that employs a DL technique. Note that the method may contain steps that do not involve DL, such as rule-based decisions. The system must receive as input one or more medical images, and it also might receive additional input, such as patient clinical history. A semi-automated system may include a human in the process, expressly, by using additional input provided by the human. We included 45 works in total.

Study exclusion. After thoroughly reading each article selected, we used two exclusion criteria to discard works that were not relevant for this survey: first, if the article did not propose a specific computational approach to solve the report generation problem, for example, if it presented a web application using existing methods or presented an assessment of feasibility, and second, if the task being addressed was different from natural language report generation from medical images, for example, report summarizing, disease classification from images, medical image segmentation, or any others. We ruled out 5 works with these exclusion criteria, leaving a total of 40 articles. The amount of articles found in each step is detailed in Table 1.

[1]In practice, most datasets reviewed present one or more pathologies, since the detection of medical conditions is one of the main motivations of these studies.

4 RESEARCH QUESTIONS

This survey aims to answer the following research questions regarding the task of *natural language report generation from medical images*:

(1) What datasets are used in this area? What diseases and imaging techniques are considered?
(2) What deep learning methods are the most commonly employed?
(3) What explainability or interpretability techniques are used?
(4) How are the proposed models evaluated? What metrics are used?
(5) How is the performance of the automatic methods? Which method can be considered *state of the art* or showing the best performance?
(6) What are the main unsolved challenges? What are the potential avenues for future work?

5 ANALYSIS OF ARTICLES REVIEWED

5.1 Datasets

We identify 18 *report datasets* containing images and reports written by experts, and 9 *classification datasets*, which provide an image and the presence or absence of a list of abnormalities. Most of the collections are publicly available (10 and 8 report and classification datasets, respectively), while the rest are proprietary. In most cases, the datasets focus on one or more pathologies and include both samples with presence and absence of these. Table 2 presents the main characteristics for the public collections, including a list of articles that used them. We next discuss the main remarks regarding report and classification datasets.

The third column in Table 2 lists the image modalities for each dataset, showing chest X-rays concentrating most of the efforts in report datasets [19, 28, 44, 69, 85], though there are also datasets with biomedical images from varied types [35, 39, 68, 105], mammography [99] and hip X-rays [37], ultrasound images [7, 158], retinal images [58], Doppler echocardiographies [98], cervical images [94], and kidney [95] and bladder biopsies [163]. This adds an extra challenge, since different kinds of exams may need different solutions, as the clinical conditions will be diverse. For example, a fundus retinal image may differ significantly from a chest X-ray, or a radiologist analyzing an X-ray may follow a different procedure than a pathologist reading a biopsy.

From the public report datasets, IU X-ray [28] is the most commonly used, consisting of 7,470 frontal and lateral chest X-rays and 3,955 reports. Additionally, each report was manually annotated with **Medical Subject Heading (MeSH)**[2] [115] and RadLex [81] terms, and automatically annotated with MeSH terms using the MTI [100] system plus the negation tool from MetaMap [10]. Figure 1 shows a sample image and report from this dataset. Note that for deep learning methods, the amount of data may seem insufficient, compared to general domain datasets with millions of samples, such as ImageNet [29]. This issue could be addressed with pre-training or data augmentation techniques. Also, this may be partially solved with the more recent datasets MIMIC-CXR [69] or PadChest [19], which contain 377,110 and 160,868 images, respectively, but have not been widely used yet.

All report datasets include images and reports, and most of them also include labels for each report. Furthermore, INbreast [99] includes contours locating the labels in the images, the Ultrasound collection [157, 158] includes bounding boxes locating organs, and IU X-ray [28] and RDIF [95] include additional text written by the physician who requested the exam. The complete details of additional information are shown in Table 9 in Appendix A.1. This information can be leveraged as a supplementary context to further improve the system performance. On the one hand, the labels and image localization can be used to design auxiliary tasks (see Section 5.2.5) and to further

[2] https://www.nlm.nih.gov/mesh/meshhome.html.

Table 2. Public Datasets Used in the Literature

Dataset	Year	Image Type	# Images	# Reports	# Patients	Used by Articles
Report datasets						
IU X-ray [28]	2015	Chest X-ray	7,470	3,955	3,955	[16, 36, 40, 48, 61, 67, 68, 85, 86, 89, 92, 120, 123, 132, 144, 150, 151, 153–156, 162]
MIMIC-CXR [69, 70]	2019	Chest X-ray	377,110	227,827	65,379	[92]
PadChest(sp) [19]	2019	Chest X-ray	160,868	109,931	67,625	None(5)
ImageCLEF Caption 2017 [35]	2017	Biomedical(2)	184,614	184,614	–	[51]
ImageCLEF Caption 2018 [39]	2018	Biomedical(2)	232,305	232,305	–	None(5)
ROCO [105]	2018	Multiple radiology(3)	81,825	81,825	–	None(5)
PEIR Gross [68]	2017	Gross lesions	7,442	7,442	–	[68]
INBreast(pt) [99]	2012	Mammography X-ray	410	115	115	[87, 128]
STARE [58]	1975	Retinal fundus	400	400	–	None(5)
RDIF(1) [95]	2019	Kidney Biopsy	1,152	144	144	[95]
Classification Datasets						
CheXpert [63]	2019	Chest X-ray	224,316	0	65,240	[156, 162]
ChestX-ray14 [143]	2017	Chest X-ray	112,120	0	30,805	[16, 67, 86, 89, 144, 151, 153]
LiTS [25]	2017	Liver CT scans	200	0	–	[131]
ACM Biomedia 2019 [55]	2019	Gastrointestinal tract (4)	14,033	0	–	[49]
DIARETDB0 [73]	2006	Retinal fundus	130	0	–	[148]
DIARETDB1 [72]	2007	Retinal fundus	89	0	–	[148]
Messidor [1, 27]	2013	Retinal fundus	1,748	0	874	[148]
DDSM [54]	2001	Mammography X-ray	10,480	0	–	[76]

All reports are written in English, except those marked with (sp) which are in Spanish, and (pt) in Portuguese. Other notes, (1): the RDIF dataset is pending release. (2): for the ImageCLEF datasets, images were extracted from PubMed Central articles and filtered automatically in order to keep only clinical images, but some unintended samples from other domains are also included. (3): contains multiple modalities, namely CT, Ultrasound, X-Ray, Fluoroscopy, PET, Mammography, MRI, Angiography and PET-CT. (4): the images are frames extracted from videos. (5): none of the articles reviewed used this dataset.

evaluate the text generation process (see Section 5.4). On the other hand, the indication may contain additional information not present in the image, such as a patient's previous condition, which in some cases may be essential to address the task [95].

Lastly, many works use classification datasets, which do not provide a report for each image, but a set of clinical conditions or abnormalities present or absent in the image. In most cases, this kind of information is used to perform image classification as pre-training, an intermediate, or an auxiliary task to generate the report. One remarkable case is the CheXpert dataset [63], which contains 224,316 images and was also presented with the CheXpert labeler, an automatic rule-based tool that annotates 14 labels (abnormalities) as present, absent, or uncertain from the natural language reports. This tool was used to label the images from the dataset and also used in MIMIC-CXR [69] to tag the reports, and in some works to evaluate the generated reports, as discussed in the Metrics section (Section 5.4). Notice that the classification dataset list is not comprehensive, as it only includes datasets that were used in at least one of the reviewed works.

Synthesis. The datasets cover multiple image modalities and body parts, though most efforts focus on chest X-rays. This opens a potential research avenue to explore other image types and diseases, using existing solutions or raising new methods. Additionally, most collections provide valuable supplementary information, such as abnormality tags and/or localization, which can be used to design auxiliary tasks and to evaluate the performance.

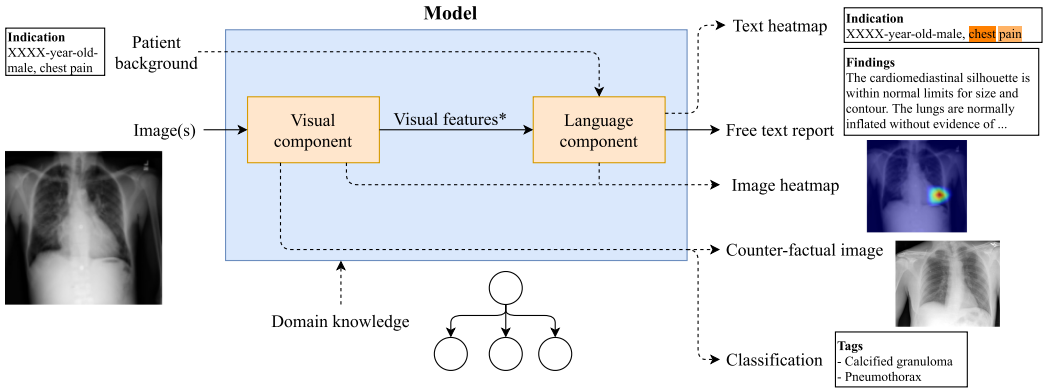


Fig. 2. General scheme of components of the architectures reviewed, including inputs on the left and outputs on the right. The blue box represents the whole model, while the orange boxes show the inner components. Solid line arrows show the flow shared by almost every work reviewed, while dashed line arrows show optional inputs and outputs seen only in some articles. Note *: in some works, the visual component may transfer classification or segmentation outputs besides or instead of *visual features*.

5.2 Model Design

In this section, we present an analysis of existent DL model designs, starting with a general overview of common design practices. Most models in the literature follow a standard design pattern. There is a visual component consisting at its core of a **Convolutional Neural Network (CNN)** [79] that processes one or more input images in order to extract visual features. Then, a language component follows, typically based on well-known NLP neural architectures (e.g., LSTM [57], BiLSTM [42], GRU [26], Transformer [139]) responsible for text processing and report generation. Also, a widespread practice for the language component is to retrieve the visual information in an adaptive manner via an attention mechanism, as the report is written. Many articles follow variations of this pattern inspired by influential works from the image captioning domain [141, 152], which are frequently cited and used as baselines. Optionally, some models receive or generate additional input or output, and a few models incorporate some form of domain knowledge explicitly in the generation process. Figure 2 presents a summary illustration of a general model architecture found in the literature. Next, we analyze model designs according to six dimensions: (1) input and output, (2) visual component, (3) language component, (4) domain knowledge, (5) auxiliary tasks, and (6) optimization strategies.

5.2.1 Input and Output. Table 3 presents a summary of this analysis.

Input. With respect to image type, most articles (24) used chest X-rays, whereas the other articles are more or less equally distributed over other image types. A total of 32 models receive a single image (e.g., a single chest X-ray view), 6 models receive two images (both frontal and lateral chest X-ray views), and 2 models receive an arbitrary number of images (e.g., multi-slice abdominal CT scans). Most models in the literature only handle visual input. However, six works [7, 16, 36, 61, 95, 132] explored the use of complementary input text, reporting performance gains in most cases. For example, two works [61, 95] encode an *indication* paragraph with a BiLSTM. Similarly, MTMA [132] encodes the report's *indication* and *findings* sections with a BiLSTM per sentence first, and then a LSTM produces a final vector representation. Similarly, two works [7, 36] use LSTM/BiLSTM to encode a partial report or caption as input, in order to predict the next word. Unlike other works, CLARA [16] uses a software package, *Lucene* [18], to perform text-based retrieval of report

Table 3. Summary of Input and Output Analysis of the Reviewed Literature

Category	Value or Type	Used by Articles
Input		
Image Type	Chest X-ray	[16, 36, 40, 44, 48, 61, 67, 68, 85, 86, 89, 92, 120, 123, 126, 132, 144, 150, 151, 153–156, 162]
	Mammography X-ray	[76, 87, 128]
	Hip X-ray	[38]
	Ultrasound video frames	[7, 76, 157, 158]
	CW Doppler echocardiography	[98]
	Gastrointestinal tract examination frames	[49]
	Gross lesions	[68]
	Bladder biopsy	[163]
	Kidney biopsy	[95]
	Liver tumor CT scans	[131]
	Cervical neoplasm WSI	[94]
	Spine MRI	[47]
	Fundus retinal images	[148]
	Biomedical images	[51]
Number of images	1	[7, 16, 36, 38, 40, 44, 47–49, 51, 61, 67, 68, 76, 87, 89, 92, 94, 98, 120, 123, 126, 128, 132, 144, 148, 151, 153, 155, 157, 158, 163]
	2	[85, 86, 150, 154, 156, 162]
	Any	[95, 131]
Text	Indication	[61, 95]
	Indication and findings	[132]
	Prefix sentence and keywords	[16]
	Partial report or caption	[7, 36]
Output		
Report	Generative multi-sentence (unstructured)	[36, 44, 48, 61, 67, 68, 89, 92, 95, 123, 128, 132, 144, 150, 151, 153–156, 162]
	Generative multi-sentence structured	[131, 163]
	Generative single sentence	[7, 38, 40, 51, 87, 120, 126, 148, 157, 158]
	Templat -based	[47, 49, 76, 94, 98]
	Hybrid template - generation/edition	[16, 85, 86]
Classification	MeSH concepts or similar	[44, 48, 49, 68, 120, 128, 132, 155, 156]
	Abnormalities/diseases presence or absence	[16, 67, 86, 89, 126, 144, 151, 157, 158, 162]
	Abnormalities/diseases characterization or severity level	[38, 76, 94, 163]
	Body parts or organs	[7, 98, 157, 158]
	Image modality	[51]
	Normal or abnormal sentence	[48, 67, 150]
Image Heatmap	Attention based per word	[92, 144, 163]
	Attention based per sentence	[61, 68, 153, 154]
	Attention based per report	[86]
	CAM [165]	[49, 94]
	Grad-CAM [118]	[89, 156]
	SmoothGrad [124]	[38]
	Activation-based attention [77]	[126]
	Bounding box (Faster R-CNN [112])	[76, 157]
	Disease and body part pixel-level classification	[47, 131]
Text Heatmap	Attention based per word	[61]
Others	Counter-factual example generation	[126]

templates. The input text is processed by *Lucene* as a search query, and the retrieved templates are paraphrased by an encoder-decoder network to generate the final report.

Output. All models output a natural language report. According to the extension of the report and the general strategy used to produce it, we group articles into five categories: (1) *generative multi-sentence (unstructured)*: these models generate a multi-sentence report, word by word, with freedom to decide the number of sentences and the words in each sentence; (2) *generative multi-sentence structured*: similar to the previous category, but always output a fixed number of sentences, and each sentence always has a pre-defined topic—these models are designed for datasets where

reports follow a rigid structure; (3) *generative single-sentence*: generate a report word by word, but only output a single sentence—these models are designed for datasets with simple one-sentence reports; (4) *template based*: use human-designed templates to produce the report, for example, performing a classification task followed by if-then rules, template selection, and template filling—this simplifies the report generation task for the model, at the expense of making it less flexible and requiring human designing of templates and rules; and lastly (5) *hybrid template - generation/edition*: use templates and also have the freedom to generate sentences word by word—this can be accomplished by choosing between a template or generating a sentence from scratch [85], or by editing/paraphrasing a previously selected template [16, 86].

In addition to the report itself, many models also output complementary classification predictions, such as presence or absence of abnormalities or diseases, MeSH concepts, and body parts or organs, among others. These are often referred to as labels or tags and are commonly used in the language component, as will be discussed in Section 5.2.3. Many models can also output heatmaps over an image highlighting relevant regions using different techniques, such as explicit visual attention weights computed during report generation, saliency map methods (e.g., CAM, Grad-CAM, SmoothGrad, or activation-based attention), bounding box regression, and pixel-level classification (image segmentation). Also, one model [61] can output a heatmap over its input text and one model [126] can generate a counter-factual example to justify its decision. We will discuss all these outputs more in detail and their use in the explainability section (Section 5.3).

5.2.2 Visual Component. The most important observation is that all surveyed works use CNNs to process the input images. This is not surprising since CNNs have dominated the state of the art in computer vision for several years [74]. The typical visual processing pipeline consists of a CNN that receives an input image and outputs a volume of feature maps of dimensions $W \times H \times C$, where W and H denote spatial dimensions (width and height) and C denotes the channel dimensions (depth or number of feature maps). These visual features are then leveraged by the language component to make decisions for report generation (e.g., which sentence to write, which template to retrieve, next word to output, etc.), typically by way of an attention mechanism.

However, some works did not strictly follow this pattern. For example, in two works [44, 128] a CNN is used for multi-label classification of tags, which are then mapped to embedded vectors via embedding matrix lookup. Thus, the report generation module only has access to these tag vectors but no access to the visual features themselves. Similarly, two works [68, 155] classify and look up tag embedding vectors, but unlike the previous works, the language component uses co-attention to access both tag vectors and visual features simultaneously. Their ablation analysis showed that the semantic information provided by these tags complements the visual information and improves the model's performance in report generation. Other works [86, 162] used graph neural networks immediately after the CNN to encode the visual information in terms of medical concepts and their relations. Thus, the language component receives the intermediate graph representation instead of the raw visual features. The ablation analysis by Zhang et al. [162] showed some performance gains thanks to the graph neural network. Vispi [89] implements a two-stage procedure, where two distinct CNNs are used. In the first stage a DenseNet 121 [60] classifies abnormalities in the image, and then Grad-CAM [118] is used to localize and crop a region of the image for each detected class. Then, in the second stage the multiple image crops are treated as independent images and processed by a typical CNN+LSTM architecture, with ResNet 101 [52] as the CNN. A similar idea was followed in RTMIC [151], where a DenseNet 121 is pretrained for classification in ChestX-ray14 [143] and CAM is used to get image crops for each class.

We observe a wide variety of CNN architectures used in the literature, though most works employ standard designs. Table 4 presents a summary. The most common ones are ResNet (11 works),

Table 4. Summary of Convolutional Neural Network Architectures Used in the Literature

Architecture	Used by Articles
DenseNet [60]	[16, 38, 85, 86, 89, 92, 151, 155, 162]
ResNet [52]	[40, 44, 48, 61, 67, 89, 94, 144, 153, 154, 156]
VGG [122]	[7, 36, 40, 51, 68, 76, 87, 95, 98, 157, 158]
Faster R-CNN [112]	[76, 157]
Inception V3 [130]	[123]
GoogLeNet [129]	[120]
MobileNet V2 [59]	[49]
SRN [166]	[44]
U-Net [116]	[128]
EcNet (*)	[163]
FCN + shallow CNN (*)	[131]
RGAN (*)	[47]
StackGAN [159] *(slightly modified version)* (*)	[126]
CNN (*)	[126, 132]
CNN *(unspecified architecture)*	[148, 150]

RGAN stands for recurrent generative adversarial network, FCN for fully convolutional network and EcNet is the name given in MDNet [163] to the custom CNN used. (*): indicates an *ad hoc* architecture.

Table 5. Summary of Language Component Architectures Used in the Literature

Architecture	Used by articles
GRU	[120]
LSTM	[40, 44, 51, 87, 120, 123, 128, 148, 157, 158]
LSTM with attention	[38, 89, 131, 144, 163]
Hierarchical LSTM with attention	[61, 68, 92, 132, 155, 156, 162]
Hierarchical: Sentence LSTM + Dual Word LSTM (normal/abnormal)	[48, 67, 150]
Recurrent BiLSTM-attention-LSTM	[95, 153, 154]
Partial report encoding + FC layer (next word)	[7, 36]
Transformer	[151]
ARAE	[126]
Template based	[47, 49, 76, 94, 98]
Hybrid template retrieval + generation/edition	[16, 85, 86]

ARAE stands for adversarially regularized autoencoder.

VGG (11 works), and DenseNet (9 works). Other standard architectures used are Faster R-CNN, Inception V3, GoogLeNet, MobileNet V2, **Spatial Regularization Network (SRN),** and U-Net. Five works used ad hoc architectures not previously published (marked with (*) in Table 4). For example, EcNet is an ad hoc architecture used in MDNet [163] and was proposed as an improvement over ResNet. However, the authors acknowledged that its design resembles DenseNet, which was published the same year (2017). RGAN, proposed by Han et al. [47], is a novel architecture that follows the **generative adversarial network (GAN)** [41] approach, with a generative module comprising the encoder and decoder parts of an **atrous convolution autoencoder (ACAE)** with a spatial LSTM between them. Similarly, Spinks and Moens [126] used a slightly modified version of a StackGAN [159] to learn the mapping from report encoding to chest X-ray images, and a custom CNN to learn the inverse mapping. Both are trained together, but only the latter is part of the report generation pipeline during inference.

5.2.3 Language Component. The job of the language component is to generate the report. In contrast to the visual component, in the literature we find a greater variety of approaches and creative ideas applied to this component. Table 5 presents a high-level summary of this analysis.

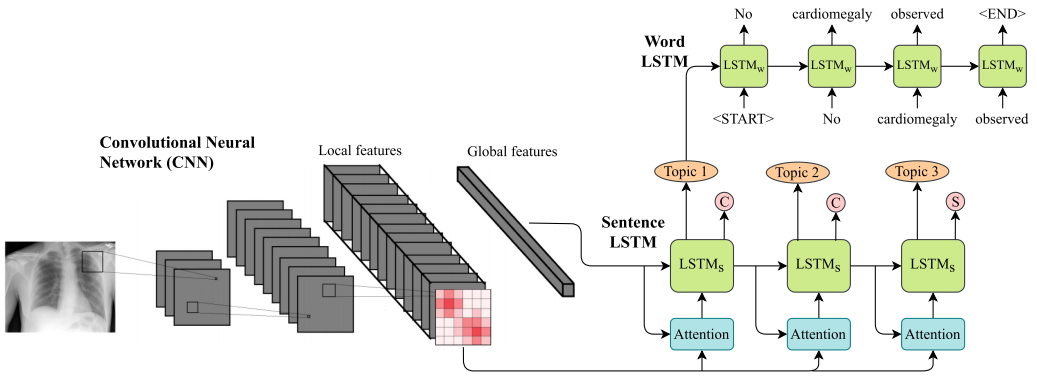


Fig. 3. Illustration of a model following the *Hierarchical LSTM with attention* approach, with attention at the sentence level. The visual component consists of a CNN. The *global features* vector can be computed from the *local features* in many ways, e.g., global average pooling. In each step the sentence LSTM generates a topic vector representing the current sentence and decides whether to stop (S) generating or continue (C).

The simplest approach is the use of a recurrent neural network, such as LSTM or GRU, to generate the full report word by word. Nine works [40, 44, 51, 87, 123, 128, 148, 157, 158] used LSTM and one work [120] tried both GRU and LSTM. All these works have in common that the GRU/LSTM receives an encoding vector from the visual component at the beginning and the full report is decoded from it. This encoding vector is typically a vector of global features output by the CNN. However, two of these works [44, 128] compute a weighted sum of tag embedding vectors and provide that as input to the LSTM. Five works [38, 89, 131, 144, 163] used LSTM enhanced with an attention mechanism. In addition to the initial input, the LSTM equipped with attention can selectively attend to visual features from the visual component at each recurrent step. This typically leads to improved performance in all articles.

A known problem for recurrent networks such as LSTM is that they are not very good at generating very long texts [103]. This is not a worrying issue when reports are short; however, it can become one for long multi-sentence reports. Two articles [131, 163] worked around this by generating each sentence independently with a single LSTM and then concatenating these sentences together. They accomplished this by providing the LSTM with a vector that indicates the sentence type as first input. This worked well in their case because the models were designed for structured reports, i.e., a fixed number of sentences per report and a fixed topic per sentence. Vispi [89] adopts a similar strategy: for each disease a dedicated LSTM generates the corresponding sentence, and the final report is the concatenation of them.

To tackle the generation of unstructured multi-sentence reports, a group of articles followed what we call the *Hierarchical LSTM with attention* approach: a Sentence LSTM generates a sequence of *topic vectors*, and a Word LSTM receives a topic vector and generates a sentence word by word. In this setting, the attention mechanism can be present at the sentence level, the word level, or both. Figure 3 shows an illustrative example. Seven works [61, 68, 92, 132, 155, 156, 162] followed this approach. A common result in these articles is that a Hierarchical LSTM yields better performance in multi-sentence report generation than a single, flat LSTM. A few articles [48, 67, 150] went one step further and replaced the normal Word LSTM with a Dual Word LSTM: the model has a gating mechanism at the sentence level that decides if the sentence will describe an abnormality (e.g., a detected cardiomegaly) or a healthy case. Thus, there are two Word LSTMs, one for normal and one for abnormal sentences. The goal is to improve the generation of abnormal sentences by having a Word LSTM that specializes in generating them. In contrast, a single Word LSTM for everything

can lead to overlearning of normal sentences and underlearning of abnormal ones, as the latter are typically less frequent due to class imbalances in datasets. The ablation analyses of these works show performance gains, thanks to this approach.

Another approach for multi-sentence report generation is the *Recurrent BiLSTM-attention-LSTM* approach. The basic idea is to have an LSTM generate one sentence at a time, each time conditioned on a BiLSTM-based encoding of the previous sentence and the output of an attention mechanism. The process is repeated recurrently sentence by sentence until the full report is generated. Three articles used this approach [95, 153, 154].

Two works [7, 36] approached report generation as simply learning to predict the next word given a partial report and an image. The models have dedicated components, such as LSTM and BiLSTM, for encoding the partial report and the image, and the next word is predicted by an FC layer. This approach simplifies the task (i.e., predict the next word given everything that comes before) but in practice requires that the model be applied recurrently one word at a time to produce a full report, which has quadratic instead of linear complexity.

Only one work, RTMIC [151], has explored the use of the Transformer [139] architecture for report generation. In RTMIC multiple image crops are obtained using Grad-CAM, then from each crop a feature vector is obtained, and finally a Transformer converts these vectors into a report. The article's results show some performance gains in CIDEr and BLEU with respect to some baselines that do not use the Transformer. Likewise, Spinks and Moens [126] were the only ones to use an **adversarially regularized autoencoder (ARAE)** [164] to generate reports. Their model combines an ARAE with a StackGAN and a normal CNN, achieving better performance than a convolutional caption generation baseline in several NLP metrics.

We also identify a group of articles [47, 49, 76, 94, 98] following a *Template-based* approach. The language component in these works operates programmatically by following if-then rules or other heuristics in order to retrieve, fill, and/or combine templates from a database in order to generate a report. The visual component typically outputs discrete classification labels that the language component processes programmatically. In the case of Harzig et al. [49], image localizations per class are also recovered using CAM [165], and in the case of Han et al. [47], the visual component outputs an image segmentation. In both cases the language component includes special localization-based rules or templates, thus incorporating location information in the generated report. Kisilev et al. [76] followed a different approach: a multi-layer perceptron learns to map image encodings to doc2vec [83] representations of corresponding reports. During inference, the ground-truth report with the closest doc2vec representation is retrieved.

Lastly, we identify three articles [16, 85, 86] following the *Hybrid template retrieval + generation/edition* approach. These works seek to combine the benefits of templates with the flexibility of a generative module to either generate sentences from scratch or paraphrase templates as needed on a case-by-case basis. KERP [86] uses ***Graph Transformers (GTRs)*** to map the visual input into a sequence of templates from a curated database. A *Paraphrase* GTR then maps each template to its paraphrased version. HRGR [85] follows the hierarchical LSTM approach with a twist—it replaces the Word LSTM with a gate module that chooses between two options: retrieving a template or generating a sentence from scratch (via a Word LSTM). Lastly, CLARA [16] is somewhat different, as it was designed as an interactive tool to assist a human to write reports. A human introduces *anchor words* and the prefix of a sentence, and *Lucene* [18] processes them as a query to retrieve sentence templates from a database. A sequence-to-sequence network then reads and paraphrases each sentence template to get the final report. CLARA can also operate fully automatically by receiving an empty prefix and predicting the anchor words itself. According to reported results, the model consistently achieved better performance than many baselines.

5.2.4 Domain Knowledge. Although all works used datasets from the medical domain to train their models, which can be considered a form of domain knowledge transfer, some works took special steps to explicitly incorporate additional knowledge from experts into their design. Concretely, we identify two incipient trends in the application of domain knowledge: (1) the use of graph neural networks right after the CNN, providing an architectural bias to guide the model to identify medical concepts and their relations from the images, and (2) enhancing the model's report generation with access to an external template database curated by experts.

KERP [86] incorporates knowledge at the architectural level using graph neural networks. The authors manually designed an abnormality graph and a disease graph, where each node represents an abnormality or disease, and the edges are built based on their co-occurrences in the training set. Some example abnormalities are "low lung volumes" and "enlarged heart size," whereas diseases represent a higher level of abstraction, for example, "emphysema" or "consolidation." The information flows from image features (encoded by a CNN) to the abnormality graph, and then to the disease graph, via inter-node message passing. This biases the network to encode the visual information in terms of abnormalities, diseases, and their relations. Similarly, Zhang et al. [162] created an observations graph, containing 20 nodes of chest abnormalities or body parts, where conditions related to the same organ or tissue are connected by edges. Their ablation analysis showed some performance gains, thanks to the graph neural network.

In seven works [16, 47, 49, 76, 85, 86, 98] the authors provided their models with a curated set of template sentences that are further processed in the language component to output a full report. Three works [47, 49, 76] used manually curated templates and if-then-based programs to select and fill them. CLARA [16] uses a database indexing all sentences from the training set reports for text-based retrieval, which are then paraphrased by a generative module. Similarly, KERP [86] has access to a template database mined from the training set, which are also paraphrased later. In HRGR [85] the most common sentences in the datasets were mined and then manually grouped by meaning to further reduce repetitions. In this work the authors showed that HRGR learned to prefer templates about 80% of the time and only generated sentences from scratch the remaining 20%, suggesting that templates can be quite useful to generate most sentences in reports.

5.2.5 Auxiliary Tasks. Although the main objective in most articles is to learn a model for report generation from medical images, many works also include and optimize auxiliary tasks to boost their performance. A summary of these tasks is presented in Table 10 in Appendix A.2. The most common auxiliary tasks are multi-label (16 papers) and single-label (11 papers) classification. These tasks are generally intended to provide additional supervision to the model's visual component, in order to improve the CNN's capabilities to extract quality visual features. Some common tasks are identifying the presence or absence of different abnormalities, diseases, organs, body parts, medical concepts, detecting image modality, and so forth. Datasets often used for this purpose are ChestX-ray14 [143] and CheXpert [63], where the common practice is to pretrain the CNN in those datasets before moving on to report generation. Many papers report better performance in report generation thanks to these auxiliary classification tasks. The three works [48, 67, 150] following the hierarchical approach with Dual Word LSTM used a classification task to supervise the gating mechanism that chooses between generating a normal sentence, generating an abnormal sentence, or stopping. Two models [47, 131] perform a segmentation task. Tian et al. [131] trained a **fully convolutional network (FCN)** with segmentation masks of a liver and tumor, and Han et al. [47] trained an RGAN for pixel-level classification. Similarly, two models [76, 157] use a Faster-RCNN [112] trained for detection and classification of bounding boxes enclosing lesions or other regions of interest in the images.

Two works [95, 155] used regularization supervision on attention weights. CORAL8 [95] receives regularization supervision on its visual attention weights to prevent them from degrading

into uniform distribution, which would offer no advantage over average pooling. Similarly, Yin et al. [155] added two regularizations to their model's attention weights: one on the weights over spatial visual features and another on the weights over tag embedding vectors. In both works the attention supervision provided a significant contribution to the performance.

Two works [98, 155] included a task to enforce a matching between embeddings from two different sources. Yin et al. [155] projected the topic vectors from the Sentence LSTM and the word embeddings from the respective ground-truth sentence into a common semantic space and enforced a matching via contrastive loss [24]. This task significantly improved the Sentence LSTM's training and the model's overall performance. Moradi et al. [98] trained an MLP for mapping image visual encodings (obtained by a VGG network) to the vector representation of its corresponding ground-truth report (obtained via doc2vec [83], which in itself was another auxiliary task), by minimizing the Euclidean distance. The trained MLP was then used to predict doc2vec representations for unseen images and retrieve the report with the closest representation. Two works [126, 132] used text autoencoders, which allow learning compact representations of unlabeled data in a self-supervised manner: an encoder network maps the input into a latent representation, and a decoder network has to recover the original input back. MTMA [132] uses a BiLSTM to encode the sentences of the *indication* and *findings* sections of a report (input text) in order to generate the *impression* section (output). To improve the encoding quality of the BiLSTM, the authors trained the decoder branch of a hierarchical autoencoder [88] to recover the original sentence from the BiLSTM encoding. The experimental results showed that the autoencoder supervision provided a significant boost to the model's performance. Spinks and Moens [126] trained an ARAE [164] (1) to learn compact representations of reports (serving as input to a GAN that generates chest X-ray images) and (2) to recover a report given an arbitrary compact representation (used in inference mode for report generation).

Lastly, Spinks and Moens [126] were the only ones to also implement cycle-consistency tasks [167] to train a GAN and an inverse mapping CNN together, to make both chest X-ray image generation and encoding more robust. These tasks will be further detailed in the next section.

5.2.6 Optimization Strategies. In addition to the architecture and the tasks a model can perform, a very important aspect is the optimization strategy used to learn the model's parameters. In this section we present an analysis of the optimization strategies used in the literature. A summary of this section is presented in Table 11 in Appendix A.3.

Visual Component. We first analyze the visual component optimization, identifying three general optimization decisions. The first one is whether to use a CNN from the literature with its weights pretrained in ImageNet [29]. This is a very common transfer learning practice from the computer vision literature in general [78], so it is natural to see it used in the medical domain too. However, it has been shown that ImageNet pretraining may not transfer as well to medical image tasks as they normally do to other domains, due to very dissimilar image distributions [109]. Therefore, a very common second decision is whether or not to train/fine-tune the visual component with auxiliary medical image tasks, such as most of the classification and segmentation tasks discussed in the previous section (Section 5.2.5). The third decision is whether to freeze the visual component weights during report generation training or continue updating them in an end-to-end manner.

Report Generation. We identify two general optimization strategies in the literature: ***Teacher-forcing* (TF)** and ***Reinforcement Learning* (RL)**. Teacher-forcing [146] is by far the most common, as it is adopted by 32 papers [7, 16, 36, 38, 40, 44, 48, 51, 61, 67, 68, 86, 87, 89, 95, 120, 123, 126, 128, 131, 132, 144, 148, 150, 153–158, 162, 163]. The basic idea in teacher-forcing is to train a model to predict each word of the report conditioned on the previous words, therefore learning to imitate the ground truth word by word. The model typically has a softmax layer that predicts the next word, and cross-entropy is the loss function of choice to measure the error and compute

gradients for backpropagation. We think teacher-forcing is so widespread in the literature because of its simplicity and general applicability, as it is agnostic to the application domain (whether it be report generation in medicine or captioning of everyday images).

In contrast, five works [67, 85, 87, 92, 151] explored the use of RL [71]. The main reason to use RL is the flexibility it offers to optimize non-differentiable reward functions, allowing researchers to be more creative and explore new rewards that may guide the model's learning toward domain-specific goals of interest. For example, Liu et al. [92] used RL to train their model to optimize the weighted sum of two rewards: (1) a natural language reward (CIDEr [140]) and (2) a ***Clinically Coherent Reward* (CCR)**, where the latter was proposed to measure the clinical accuracy of a generated report compared to a ground-truth reference using the CheXpert labeler tool [63]. Their goal was to equip their model with two skills: natural language fluency (encouraged by CIDEr) and clinical accuracy (encouraged by CCR). Other examples of the use of RL are the direct optimization of CIDEr [85, 151], particularly in the training of a complicated hybrid template retrieval and text generation model [85]; directly optimizing BLEU-4 after a previous teacher-forcing warmup phase [67]; and the training of the generator network of a GAN used for report generation, where the reward is provided by the discriminator network [87].

As a side note, we would like to highlight the work by Zhang et al. [161] on medical report summarization (a related task where the report is the input and with no images), illustrating how RL can be used in this setting to optimize both fluency and factual correctness. As rewards they used ROUGE [90] and a Factual Correctness reward based on the CheXpert labeler tool [63] (very similar to the CCR proposed by Liu et al. [92]). This work is a good example of the benefits of RL over teacher-forcing for text generation in a medical domain. The paper presents the results of a human evaluation with two board-certified radiologists, and the model trained with RL achieved better results than the same model trained with teacher-forcing, and even slightly better results than the human baseline.

Other Losses or Training Strategies. This category encompasses the remaining optimization strategies found in the literature. The most important one is *multitask learning* [21], adopted by 14 papers [48, 67, 68, 76, 86, 94, 95, 126, 131, 132, 144, 155, 157, 163]. The main idea is to jointly train a model in multiple complementary tasks, so that the model can learn robust parameters that perform well in all of them. Some works [48, 68, 131, 132, 144, 155, 163] trained the visual and language components simultaneously in multiple tasks in an end-to-end manner, i.e., report generation plus other auxiliary tasks. Other examples are the simultaneous training of object detection and attribute classification [76], diagnostic classification and cycle-consistency tasks [126], among others. Most of these papers report benefits from training in this way.

As already discussed in Section 5.2.5, two works [95, 155] used auxiliary supervision on the attention weights of their models. These auxiliary losses were jointly optimized with the rest of the model in report generation, effectively having a regularizer effect. Yin et al. [155] are also the only ones that included an auxiliary contrastive loss [24] to provide a direct supervision to the Sentence LSTM, thus improving their model's performance. Notice that all these works are examples of multitask learning too. Three papers [76, 98, 157] used regression losses. Two of them [76, 157] included a bounding box regression loss as part of Faster R-CNN [112] training, and Moradi et al. [98] included a regression loss to minimize the Euclidean distance between VGG and doc2vec embeddings. As previously discussed in Section 5.2.5, another optimization strategy is the use of autoencoders for the self-supervised learning of text representations. In MTMA [132] an autoencoder was used to provide an auxiliary supervision over the BiLSTM and was jointly trained with the rest of the model in a multitask learning fashion. Spinks and Moens [126] instead trained an ARAE in a first stage, then froze its weights and used the learned text embedding to support the subsequent training of a GAN.

Lastly, three works used GANs [47, 87, 126]. As mentioned when discussing RL, Li and Hong [87] used a GAN strategy to train their model for report generation, where the generative module generates a report and the discriminator determines whether it is real or fake. Similarly, Han et al. [47] proposed RGAN, where the generator outputs segmentation maps from spine radiographs and the discriminator determines if a given segmentation map is real or fake. Spinks and Moens [126] implemented a modified version of a StackGAN [159] to generate chest X-ray images from input text representations. In their case, they trained the GAN using two cycle consistency [167] losses: (1) image → embedding → image and (2) embedding → image → embedding. In both cases, an auxiliary inverse mapping CNN was used to close the cycle.

Synthesis. Overall, we can observe that designing a model for report generation from medical images is a complex task that involves engineering decisions at multiple levels: inputs and outputs, visual component, language component, domain knowledge, auxiliary tasks, and optimization strategies. In each of these dimensions there are different approaches adopted in the reviewed literature, and the current state of research does not allow us to recommend an "optimal model design," mainly for reasons we will discuss in the Metrics and Performance Comparison sections (Sections 5.4 and 5.5). Nevertheless, there are valuable insights in the literature that may lead to better results, and thus are worth having in mind, for example, the use of CNNs (such as DenseNet or ResNet) as visual component and training in auxiliary medical image tasks, the use of input text alongside the images; providing the language component with tag information in addition to the visual features (e.g., medical concepts identified in the image), leveraging template databases curated with domain knowledge, or the use of multitask learning combining multiple sources of supervision. Lastly, to improve report quality from a medical perspective, the use of reinforcement learning with adequate reward functions appears as the most promising approach.

5.3 Explainability

There have been multiple attempts on providing a definition for *explainability* in the **Explainable Artificial Intelligence (XAI)** area [33, 91, 113]. For the task of report generation from medical images, we use a similar definition by Doshi-Velez and Kim [33]: *the ability to justify an outcome in understandable terms for a human*, and we use it interchangeably with the term *interpretability*. In this medical context, an automated system requires high explainability levels as two main facts hold: the decisions derived from the system will probably have direct consequences for patients, and the diagnosis task is not trivial and susceptible to human judgment [33, 113]. Furthermore, the explanation methods employed in this medical task should attempt to solve several related aspects: align with clinicians' expectations and acquire their trust, increase system transparency, assess results quality, and allow addressing accountability, fairness, and ethical concerns [4, 113, 134].

There are many ways to address the explainability aspect of AI systems in the medical domain, as listed in the recent survey on interpretable AI for radiology by Reyes et al. [113]. Multiple categories can be identified, starting with *global vs. local*; the former refers to explanations regarding the whole system's operation, and the latter to explanations for one sample. For local explanations, there are different kinds of approaches, such as *feature importance*, *concept based*, *example based*, and *uncertainty*, to mention a few. *Feature importance* methods attempt to compute a level of importance for each input value, to understand which characteristics were most relevant to make a decision, for example, gradient-based methods for CNNs such as Grad-CAM [118], Guided Backpropagation [127], or DeepLIFT [121], and other techniques such as LIME [114]. In *concept-based* methods, like TCAV [75] or RCV [43], the contributions to the prediction from multiple concepts are quantified, so the user can check if the concepts used by the model are correct. *Example-based* approaches present additional examples with the output, either with a similar outcome, so the

user can look for a common pattern, or with an opposite outcome (counter-factual). *Uncertainty* methods provide the level of confidence of the model for a given prediction. For global explanations, there are *sample-based* approaches, such as SP-LIME [114], or methods to directly increase the *transparency* of the system.

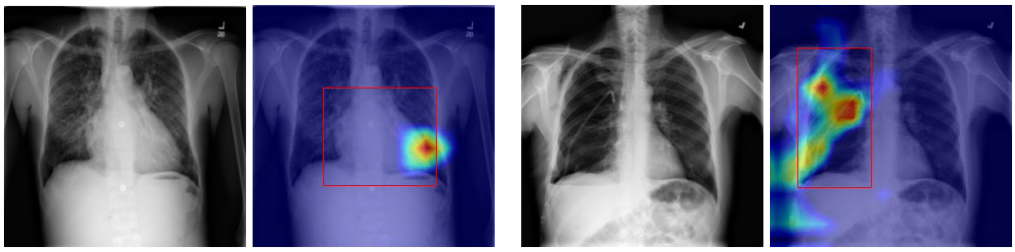
Despite the importance of explainability in this area, only two reviewed works focused explicitly on this topic. Gale et al. [38] proposed the automatic generation of a natural language report as an explanation for a classification task; however, their approach does not include an explanation for the report. Spinks and Moens [126] present a counter-factual local explanation, as will be detailed in Section 5.3.1. Additionally, in 29 works the model architecture generates a secondary output that can also be presented as a local explanation. We distinguish three types of outputs: classification (Section 5.3.2), heatmap over the input image (Section 5.3.3), and heatmap over the input text (Section 5.3.4). These were already summarized in Table 3 in the Input and Output section (Section 5.2.1). Next, the explanaibility aspects of the outputs are discussed.

5.3.1 Counter-actual. Spinks and Moens [126] proposed an architecture to both classify a disease and generate a caption from a chest X-ray, based on GANs and autoencoders, as detailed in the Model Design section (5.2). Thus, to provide a local explanation, at inference time the input image is encoded into a latent vector, which is used to generate a new chest X-ray and a new report, both of them subject to result in the nearest alternative classification, i.e., the nearest diagnosis. With this information, a user could compare the original X-ray with the generated image and attempt to understand why the model has reached its decision.

5.3.2 Classification. As explained in the Auxiliary Tasks section (5.2.5), many deep learning architectures include multi-label classification to improve performance, providing a set of classified concepts as secondary output. Even though in most papers this kind of output is not presented as an explanation of the report, we consider that its nature could improve the transparency of the model, which is an important way of improving the interpretability in a medical context [134]. By providing this detection information from an intermediate step of the model's process, an expert could further understand the internal process, validate the decision with their domain knowledge, and calibrate their trust in the system.

As shown in Table 3 from Section 5.2.1, the terms classified are very diverse. Some works classify very broad concepts, such as body parts or organs [7, 98, 157, 158] or image modality [51]. Other works perform a more specific classification, such as diseases or abnormalities [16, 76, 86, 126, 144, 157, 158, 162] or a normal or abnormal status at sentence level [150]. Lastly, several works [44, 48, 68, 120, 128, 132, 155, 156] classify over a subset of MeSH terms or similar, which may contain a mix of general broad medical concepts and specific abnormalities or conditions. We believe that this additional output would be useful for an expert, though the specific concepts should provide much richer information. If the classification is more specific, the user will be able to validate on a much narrower scope the system's performance.

5.3.3 Image Heatmap. In the papers reviewed, there are three different approaches to generating heatmaps over the input image, each of them with a different interpretation. First, many architectures employ an attention mechanism over the image spatial features during the report generation, as was discussed in the Language Component section (Section 5.2.3). These mechanisms can be leveraged to produce a heatmap indicating the image regions that were most important to generate the report. In particular, some models provide a heatmap for each word [92, 144, 163], for each sentence [61, 68, 153, 154], or for the whole report [86]. By showing these feature importance maps, an expert should be able to determine if the model is focusing on the correct regions of the image, which could improve their trust on the system.



Cardiomegaly Pneumothorax

Fig. 4. Examples from the ChestX-ray14 dataset [143] classified with a CNN based on ResNet-50 [52] and using CAM [165] to provide a heatmap indicating the spatial regions of most importance as local explanation. The left example presents Cardiomegaly and the right Pneumothorax, and both samples were correctly classified by the CNN. Red boxes represent a localization of the condition annotated by an expert.

Second, some works use particular deep learning architectures to perform image segmentation, i.e., classification and localization at the same time. The model by Ma et al. [94] uses a CNN to classify the severity of four different key characteristics of cervical cancer and then uses an attention mechanism over the visual spatial features to generate heatmaps indicating the position of each relevant property. Tian et al. [131] used an FCN to classify each pixel of an image with the presence of a liver or tumor, and the result is averaged with an attention map to further improve localization. Han et al. [47] proposed the ACAE module (see Section 5.2.3 for details), which is used to classify at pixel level different parts of the spine (vertebrae, discs, or neural foramina) and if they show an abnormality or not. Kisilev et al. [76] and Spinks and Moens [126] used a Faster R-CNN [112] architecture to detect image regions with lesion and body parts of interest.

Lastly, some works use gradient- or activation-based methods for CNNs to generate a saliency map indicating the regions of most importance for a classification, such as CAM [165], Grad-CAM [118], SmoothGrad [124], or the one proposed by Komodakis and Zagoruyko [77]. Refer to Table 3 in the Input and Output section (5.2.1) for a list of the papers using each technique. To determine which of these methods performs better in a general setting, Adebayo et al. [3] performed multiple evaluations ("sanity checks") over Grad-CAM, SmoothGrad, and other similar methods and showed that Grad-CAM should be more reliable in terms of correlation with the input images and the classification made. As an example of these techniques, Figure 4 shows two chest X-rays from the ChestX-ray14 dataset [143] with a heatmap generated with CAM, plus an expert-annotated bounding box locating the abnormality (provided with the dataset).

In both segmentation and saliency map methods, the heatmap information provides much richer information than classification alone, as it also includes the location of a specific concept, such as an abnormality or a body part. Providing this type of explanation should allow an expert to assess the localization capabilities of the model and the system accuracy, thus improving the model's transparency throughout its process.

5.3.4 Text Heatmap. The model proposed by Huang et al. [61] also receives text as input, which indicates the reason for performing the imaging study on the patient. In a similar fashion to the input image cases, the architecture includes an attention mechanism over the input text, which provides a heatmap indicating the input phrases or sentences that were most relevant to generate each word in the output. With this feature importance map, an expert should be able to determine if the model is focusing on the correct words in the input text.

Synthesis. All the explainability approaches are local explanations given by a secondary output, either indicating feature importance (image and text heatmap), increasing the model's

transparency (classification), or providing a counter-factual example. However, in most of the works the authors do not explicitly mention it as an interpretability improvement, and in almost all cases there is no formal evaluation, as will be discussed in Section 5.4.3. Hence, we believe this is an understudied aspect of the medical report generation task, given the superficial or nonexistent analysis it receives in most of the reviewed works. Additionally, counter-factual techniques could be further studied, and other approaches not found in the literature could be explored, such as prediction uncertainty or global explanations, which may be quite relevant for clinicians [134].

5.4 Evaluation Metrics

There are different ways to assess a medical report generated by an automated system. We divide the evaluation metrics used in the literature into three categories, depending on the aspect being assessed: text quality, medical correctness, and explainability. Also, each evaluation method can be either automatic or performed manually by humans. Each of the categories and metrics are presented next, and Table 6 shows a summary of the metrics used by each paper.

5.4.1 Text Quality Metrics. The methods in this category measure general quality aspects of the generated text and are originated from translation, summarizing, or captioning tasks. The most widely used metrics in the papers reviewed are BLEU [102], ROUGE-L [90], METEOR [12, 82], and CIDEr [140], which measure the similarity of a target text (also referred to as candidate) against one or more reference texts (ground truth). These metrics are mainly based on counting n-gram matchings between the candidate and the ground truth. BLEU is precision oriented, ROUGE-L and METEOR are F1-scores that can be biased toward precision or recall with a given parameter, and CIDEr attempts to capture both precision and recall through a TF-IDF score. Most of these metrics have variants and parameters for their calculation: ROUGE is a set of multiple metrics, with ROUGE-L being the only one used in this task; METEOR has variants presented by the same authors [30–32]; and CIDEr was presented with the CIDEr-D variant to prevent gameability effects.

SPICE [9] is a metric designed for the image captioning task and evaluates the underlying meaning of the sentences describing the image scene, partially disregarding fluency or grammatical aspects. Specifically, the text is parsed as a graph, capturing the objects, their described characteristics, and relations, which are then measured against the ground truth using an F1-score. Even though SPICE attempts to assess the semantic information in a caption, we believe it is not suitable for medical reports, as the graph parsing is designed for general domain objects. Nonetheless, Zhang et al. [162] presented the medical correctness metric MIRQI, applying a similar idea in a specific medical domain, which we will discuss in the next subsection (Section 5.4.2).

Besides standard captioning metrics, we identified two other approaches to measure text quality. First, Alsharid et al. [7] used Grammar Bot,[3] a rule- and statistics-based automated system that counts the grammatical errors in sentences. Second, Harzig et al. [48] measured the sentence variability by counting the different sentences in a set of reports. They argue that the sentences indicating abnormalities occur very rarely in the dataset, while the ones indicating normality are the most frequent. Hence, a certain level of variability is desired, and a system generating reports with low variability may indicate that not all medical conditions are being captured.

Lastly, both works from Li et al. [85, 86] performed human evaluation with non-expert users via **Amazon Mechanical Turk (AMT)** following the same procedure. The authors presented two reports to the AMT participants, one generated with the proposed model and one generated with the CoAtt model [68] as baseline, and asked them to choose the most similar with the ground truth in terms of fluency, abnormalities correctness, and content coverage. The results show that their

[3] https://www.grammarbot.io/.

Table 6. Summary of the Evaluation Metrics Used in the Literature

Category	Metric or Evaluation	Used by Papers
Text quality (automatic)	BLEU based	[7, 16, 36, 38, 40, 44, 48, 51, 61, 67, 68, 85–87, 89, 92, 95, 120, 123, 126, 128, 131, 132, 144, 150, 151, 153–158, 162, 163]
	ROUGE-L	[7, 36, 44, 48, 61, 67, 68, 85, 86, 89, 92, 95, 123, 126, 131, 132, 144, 150, 153–158, 162, 163]
	METEOR based	[36, 44, 48, 68, 87, 95, 123, 126, 144, 153–158, 163]
	CIDEr based	[16, 36, 48, 61, 67, 68, 85, 86, 89, 92, 123, 126, 128, 150, 151, 153, 155, 157, 158, 162, 163]
	SPICE	[87]
	Grammar Bot	[7]
	Sentence variability	[48]
Text quality (with humans)	AMT study	[85, 86]
Medical correctness (automatic, report based)	MIRQI (precision, recall, F1)	[162]
	MeSH accuracy	[61]
	Keyword ratio (accuracy, sensitivity, specificity)	[148]
	Keyword accuracy	[150, 154]
	Medical abnormality terminology detection (precision, FPR)	[85]
	Abnormality detection (precision, FPR)	[67]
	Medical abnormality detection (accuracy, precision, recall)	[92]
	Abnormality CNN classifier (accuracy, PR-AUC)	[16]
	Semantic descriptors	[98]
	ARS	[7]
Medical correctness (automatic, auxiliary tasks)	ROC-AUC	[86, 89, 144, 162]
	Accuracy	[76, 94, 120, 157, 158, 163]
	Recall/sensitivity	[49, 155]
	Precision	[49, 76, 155]
	Specificity	[49]
	Pixel-level accuracy	[47]
	Pixel-level specificity	[47]
	Pixel-level sensitivity	[47]
	Pixel-level dice score	[47, 131]
Medical correctness (with experts)	Assess correctness of the nature of hip fractures	[38]
	Accept/reject rating	[131]
	Assess medical and grammatical correctness and relevance	[7]
	Agree with diagnosis	[126]
Explainability (with experts)	Counter-factual X-ray vs. saliency map	[126]
	Reports vs. SmoothGrad (classification explanation)	[38]

The report based medical correctness type includes metrics that are measured from the report generated; the auxiliary task medical correctness ones evaluate an auxiliary or intermediate task in the process, such as classification or segmentation.

report was preferred in around 50% to 60% of the cases, while the baseline around 20% to 30% (for the rest, none or both were preferred). We categorize this evaluation as a text quality metric, as the participants are not experts, and their answers are not fine-grained (i.e., did not specify what failed: fluency, correctness, or coverage, or by how much they failed).

5.4.2 Medical Correctness Metrics. While the most common purpose of the text quality metrics is to measure the similarity between the generated report and a ground truth, they do not necessarily capture the medical facts in the reports [11, 17, 92, 107, 108, 161]. For example, the

sentences *"effusion is observed"* and *"effusion is not observed"* are very similar and thus may present a very high score for any metric based on n-gram matching, though the medical facts are the exact opposite. Therefore, an evaluation directly measuring the reports' correctness is required, not necessarily taking into account fluency, grammatical rules, or text quality in general. From the literature reviewed, in 10 works [7, 16, 61, 67, 85, 92, 98, 148, 154, 162] the authors presented an automatic metric to address this issue, 4 works [7, 38, 126, 131] did a formal expert evaluation, and multiple works [49, 76, 86, 89, 94, 120, 126, 131, 144, 155–158, 162, 163] evaluated medical correctness indirectly from auxiliary tasks. The methods are listed in Table 6 and are further discussed next.

In several works the authors presented a method that detects concepts in the generated and ground-truth reports and compare the results using common classification metrics, such as accuracy, F1-score, and more. The main difference between these methods lies in how the concepts are automatically detected in the reports. The simplest approaches are *keyword based*, which consists in reporting the ratio of a set of keywords found between the generated report and ground truth, like MeSH Accuracy [61] that uses MeSH terms, and Keyword Accuracy that uses 438 MTI terms (presented by Xue et al. [154] and used in A3FN [150]). Similarly, Medical Abnormality Terminology Detection [85] calculates precision and false-positive rate of the 10 most frequent abnormality-related terms in the dataset, and Wu et al. [148] calculated accuracy, sensitivity, and specificity for a set of keywords.

Other approaches are *abnormality based*, which attempt to directly classify abnormalities from the report by different means: Abnormality Detection [67] uses manually designed patterns; Medical Abnormality Detection [92] uses the CheXpert labeler tool [63]; Biswal et al. [16] used a character-level CNN [160] that classifies multiple CheXpert labels [63]; and Moradi et al. [98] used a proprietary software to extract semantic descriptors. Lastly, the **Anatomical Relevance Score (ARS)** [7] is a *body-part-based* approach, which detects the anatomical elements mentioned in a report considering the vocabulary used. Though these methods may be useful for measuring medical correctness to a certain degree, there is no consensus or standard, and there is no formal evaluation of the correlation with expert judgment. From the discussed techniques, Alsharid et al. [7] are the only authors who also performed an expert evaluation of the generated reports, though they did not conduct a correlation or similar analysis to validate the ARS method.

Zhang et al. [162] went further with the concept extraction and presented the **Medical Image Report Quality Index (MIRQI)**, which works in a similar fashion as the SPICE [9] metric presented in the text quality subsection (Section 5.4.1). MIRQI applies ideas from NegBio [106] and the CheXpert labeler [63] to identify diseases or medical conditions in the reports, considering synonyms and negations, and uses the Stanford parser [23] to obtain semantic dependencies and finer-grained attributes from each sentence, such as severity, size, shape, body parts, and so forth. With this information, an abnormality graph is built for each report, where each node is a disease with its attributes, and the nodes are connected if they belong to the same organ or tissue. Lastly, the graphs from the ground truth and generated reports are matched node-wise, and MIRQI-p (precision), MIRQI-r (recall), and MIRQI-F1 (F1-score) are computed. Compared to the formerly discussed correctness metrics, we believe this approach seems more robust to assess the medical facts in the reports, as it attempts to capture the attributes and relations, as opposed to the concepts only. However, the authors did not present an evaluation against expert judgment, so we cannot determine if this metric is sufficient.

Considering human evaluation, only a few works [7, 38, 126, 131] present a formal expert medical correctness assessment. In the work by Alsharid et al. [7] a medical professional assessed the reports on a Likert Scale from 0 to 2 in four different aspects: *accurately describes the image*, *presents no incorrect information*, *is grammatically correct,* and *is relevant for the image*; the results

were further separated for samples from different body parts, showing averages between 0.5 and 1. Gale et al. [38] asked a radiologist to evaluate the correctness of the hip fractures description, finding that the fracture's character was properly described in 98% of the cases, while the fracture location only in 90%. In the work by Tian et al. [131] a medical expert evaluated 30 randomly selected reports with a rating from 1 (*definite accept*) to 5 (*definite reject*), scoring an average of 2.33. Lastly, Spinks and Moens [126] asked four questions to three experts regarding the generated reports, where the third and fourth questions measured correctness: *"Do you agree with the proposed diagnosis?,"* answering 0 (*no*) or 1 (*yes*), and *"How certain are you about your final diagnosis?,"* from 1 (*not sure*) to 4 (*very sure*). The average scores were high (0.88 and 3.75), showing agreement with the model's diagnosis, and certainty on the experts' diagnoses. The other questions concerned explainability aspects and are detailed in the next subsection (Section 5.4.3). So far, there is no standard approach to perform an expert evaluation, though we believe the first two approaches provide finer-grained information than the latter two, and hence should be more useful for determining in which cases the models are failing and for designing improvements. The certainty question should also be very useful, as diagnoses may be susceptible to human judgment.

Lastly, multiple papers [49, 86, 89, 94, 120, 131, 144, 155, 156, 158, 162, 163] evaluated the performance of the auxiliary tasks with ROC-AUC, accuracy, and other typical classification or segmentation metrics, as shown in Table 6. Note that in any of these cases, the task is a previous or intermediary step of the process and is not derived from the report. In consequence, even if the classification has great performance, the language component could be performing poorly, and the generated reports still may be inaccurate. Accordingly, we believe this type of measure should not be used as the primary report correctness evaluation, unless it can be proven that the report reproduces exactly the classification made (e.g., by a template-filling process).

5.4.3 Explainability Metrics. Providing interpretable justifications for the model's outcome is essential in this medical domain, and furthermore, we should be able to evaluate them to answer questions such as: Does the method justify the model's decision? Which method provides a *better* explanation? However, there is no consensus on evaluation methods for AI explainability, and in many cases the definition of a *better* explanation remains subjective [22, 33, 113].

Consequently, none of the papers reviewed used an automatic metric to assess explainability, and only two works [38, 126] conduct a formal human expert evaluation. Gale et al. [38] presented the report generation as an explanation of a medical image classification task and evaluated it by comparing three methods: (a) SmoothGrad [124] to highlight the most important pixels used, (b) a generated report in natural language, and (c) both placed side by side. Five experts assessed 30 images, rating each explanation in a scale from 1 (*unsatisfactory*) to 10 (*perfect*);, achieving average scores of (a) 4.4, (b) 7, and (c) 8.8 for each method. Though the authors emphasize the importance of the natural language explanations, their approach does not include an explanation for the report itself, so it cannot be directly used for the report generation task.

The model proposed by Spinks and Moens [126] generates a chest X-ray as a counter-factual example, and they compared this explanation method against a feature importance heatmap generated with the Zagoruyko and Komodakis saliency map technique [77]. Three experts evaluated 150 samples answering four questions, the first two regarding explainability aspects: *"Does the explanation justify the diagnosis?" "Does the model appear to understand the important parts of the X-ray?"* The answers were on a scale from 1 (*no*) to 4 (*yes*), and their method achieved a higher score than the saliency map (2.39 vs. 1.31 for the first question, and 2.45 vs. 1.81 for the second), showing their counter-factual approach should be *better* in this setting. The other two questions relate more to medical correctness and are discussed in the previous section (Section 5.4.2).

We believe the explanation evaluations should be very important in this area, and as there is no consensus, we outline some possible guidelines. Following ideas from Tonekaboni et al. [134], we believe three aspects from the explanations should be assessed: (1) consistency, (2) alignment with domain knowledge, and (3) user impact. First, the consistency across the data should be assessed, answering questions such as: Do explanations change with variations to the input data? Or to the prediction? Or to the model design? Or with different images from the same patient? As pointed out by Tonekaboni et al. [134], inconsistent explanations may negatively affect the clinicians' trust, and an interpretability method laying them out should be reviewed. Examples of consistency or robustness evaluations can be found in the work by Adebayo et al. [3] for image saliency maps, and in the work by Jain and Wallace [66] for attention in recurrent neural networks.

Second, the alignment with domain knowledge should evaluate if the explanation is consistent with an expert's knowledge: would they provide the same explanation for that decision? For instance, given a feature importance method, is the model focusing on the correct features? As an example, consider the second and third questions employed by Spinks and Moens [126] detailed earlier. To mention other examples, Wang et al. [143] evaluated CAM [165]-generated heatmaps for disease classification against expert-provided bounding boxes locating the diseases, using *intersection-over-union*-like metrics; Kim et al. [75] proposed a model to classify Diabetic Retinopathy from retina fundus images, and they compared the TCAV [75]-extracted concepts against expert knowledge. Notice many works reviewed in this survey used classification or segmentation as an auxiliary task, which can be used as local explanations, and evaluated them with common metrics (such as accuracy, precision, etc.), as discussed in the previous sections (Sections 5.3 and 5.4.2). As the authors did not mention the secondary outputs as local explanations, we categorized the said evaluations as *medical correctness* metrics, but they are also measuring *alignment with domain knowledge* for the interpretability methods and as such may be very useful.

Lastly, the user impact should attempt to answer questions such as: Is it a *good* explanation? Does it provide *useful* or *novel* information? Does it justify the model's decision? Is it provided with an appropriate representation for the experts? As examples, the assessment proposed by Gale et al. [38] and the first question used by Spinks and Moens [126] measure user impact. Notice that most of these concepts are very subjective, and the definitions, questions, and assessments will vary for different sub-domains and target experts. We believe more specific definitions and fine-grained aspects should arise in the future, as research in this topic grows. For reference, this category includes the *domain-appropriate representation* and *potential actionability* concepts presented by Tonekaboni et al. [134].

Synthesis. Almost all the works include text quality metrics, though these are not able to capture the medical facts in a report [11, 17, 92, 107, 108, 161]. Several works proposed medical correctness assessments over the reports, but unfortunately none of the proposals was evaluated against expert judgment. The auxiliary tasks can be evaluated to measure correctness indirectly from the process, but often it will not be sufficient for the report's correctness. Only two works evaluate explainability directly with experts, and the auxiliary tasks' assessments could be useful to measure alignment between the explanations and domain knowledge. Overall, we believe that medical correctness should be the primary aspect to evaluate in the generated reports, using one or more automatic metrics. For now, and even though none of the metrics proposed has been evaluated against expert judgment, MIRQI [162] seems like the most promising approach to fulfill this purpose, as it should be able to capture richer information from the reports. Additionally, text quality metrics can be used as a secondary evaluation, since they may be useful for measuring fluency, grammar, or variability, and to compare with previous baselines. Lastly, explainability evaluation methods should arise to assess multiple key aspects, such as its consistency, alignment with domain knowledge, and the user impact.

5.5 Comparison of Papers' Performance

To find out which paper holds the state of the art, we need to find a common ground for fair comparison. A natural choice is the IU X-ray dataset [28], since a majority of the surveyed papers report results in this dataset. Table 7 shows these results, separated by which report sections are generated by each paper, *findings*, *impression,* or both. The findings section consists of multiple sentences and mainly describes medical conditions observed, while the impression section is a one-sentence conclusion or diagnosis. Notice Spinks and Moens [126] filtered the *findings* section and kept only sentences referring to one disease (Cardiomegaly). The papers that seem to show the best performance in terms of NLP metrics are KERP [86], CLARA [16], and Xue and Huang [153] for the findings section; MTMA [132] for the impression section; and Yuan et al. [156], MLMA [36], and Xue and Huang [153] for both sections. Of these, only MTMA has a large difference to its competitors, and there is no clear winner in the other sections. Some caveats, however, should be kept in mind when interpreting these results:

(1) The results reported in the literature only allow comparisons in terms of standard natural language metrics (BLEU, ROUGE-L, etc.), but from these results we cannot draw conclusions about medical correctness, since NLP metrics and clinical accuracy are not necessarily correlated.

(2) MTMA uses additional input, as discussed in Section 5.2.1. Specifically, the model receives the indication and findings sections of the report to generate the impression section, at both test and train stages. In a sense, this could be seen as an enhanced summarizing approach, since the impression section contains a conclusion from the findings.

(3) Some NLP metrics, such as CIDEr, ROUGE-L, and METEOR, have variants and parameters, as discussed in Section 5.4. Unfortunately, most papers do not mention the specific version or implementation used.

(4) The IU X-ray dataset does not have standard training-validation-test splits. This has led researchers to define their own splits, as indicated by column *Split* of Table 7. These splits are not consistent across papers, making results less comparable. For example, if a model was evaluated in an easier test split, that would give it an unfair advantage over other models evaluated in harder test splits. Additionally, other decisions such the number of images per report (frontal, lateral, or both), the tokenization algorithm employed, the removal of noisy sentences, the removal of words with a frequency under a given threshold, and the removal of duplicate images, among other preprocessing decisions, are not always explicitly stated in papers, and these may have an impact on the results as well.

(5) These are overall results only, so a more fine-grained performance assessment on specific abnormalities or diseases is missing. This further shows the need for standardizing one or more evaluation metrics to measure the medical correctness of a generated report, considering different aspects of interest.

6 CHALLENGES AND FUTURE WORK

In this section, we identify unsolved challenges in the literature and potential avenues for future research in the task of report generation from medical images.

Protocol for expert evaluation. If the ultimate goal is to develop a report generation system that meets high-quality standards, it makes sense that such a system be thoroughly tested by medical experts to evaluate its performance in different clinical settings. Most papers reviewed are weak in this regard, as only four of them [7, 38, 126, 131] perform a correctness evaluation with medical experts, meaning that 90% of the works do not carry out an expert evaluation, feedback that should be immensely valuable to understand the strengths and weaknesses of a model. Therefore, a clear avenue for improvement is to standardize a protocol for human evaluation of these systems by

Table 7. Evaluation Results of Papers That Use the IU X-ray Dataset

Paper	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr-D	Split
Findings Section								
Liu et al. [92]	0.369	0.246	0.171	0.115	0.359	–	**1.490**	7:1:2 [1]
HRGR [85]	0.438	0.298	0.208	0.151	0.369	–	0.343	7:1:2 [2]
KERP [86]	**0.482**	**0.325**	**0.226**	0.162	0.339	–	0.280	7:1:2 [2]
TieNet [144] [(1)]	0.330	0.194	0.124	0.081	0.311	–	1.334	7:1:2 [1]
Xue et al. [154] [(2)]	0.441	0.320	0.231	0.181	0.366	0.220	0.343	2,525/250
RTMIC [151]	0.350	0.234	0.143	0.096	–	–	0.323	7:2:1
CLARA [16] [(3)]	0.471	0.324	0.214	**0.199**	–	–	0.359	7:1:2
Xue and Huang [153]	0.477	0.332	0.243	0.189	**0.380**	**0.223**	0.320	3,031/300
CMAS [67]	0.464	0.301	0.210	0.154	0.362	–	0.275	Unk
Findings Section - Cardiomegaly Sentences Only								
Spinks and Moens [126]	0.490	0.350	0.250	0.180	0.400	0.270	0.600	8:1:1
Impression Section								
MTMA [132]	**0.882**	**0.874**	**0.867**	**0.860**	**0.929**	–	–	5,461/500/500
CMAS [67]	0.401	0.290	0.220	0.166	0.521	–	1.457	Unk
Findings + Impression Sections								
CoAtt [68]	0.517	**0.386**	0.306	0.247	0.447	0.217	0.327	6,470/500/500
Huang et al. [61]	0.476	0.340	0.238	0.169	0.347	–	0.297	8:1:1
Yuan et al. [156]	**0.529**	0.372	0.315	0.255	0.453	**0.343**	–	8:2
Xue et al. [154]	0.464	0.358	0.270	0.195	0.366	0.274	–	2,775/250
Vispi [89]	0.419	0.280	0.201	0.150	0.371	–	0.553	7:1:2
Singh et al. [123]	0.374	0.224	0.153	0.110	0.308	0.164	0.360	6,718/350/350
Yin et al. [155]	0.445	0.292	0.201	0.154	0.344	0.175	0.342	6,470/500/500
MLMA [36]	0.500	0.380	**0.317**	**0.278**	0.440	0.281	**1.067**	6,429/500/500
Harzig et al. [48]	0.373	0.246	0.175	0.126	0.315	0.163	0.359	90:5:5
A3FN [150]	0.443	0.337	0.236	0.181	0.347	–	0.374	9:1
Xue and Huang [153]	0.489	0.340	0.252	0.195	**0.478**	0.230	0.565	3,031/300
Zhang et al. [162]	0.441	0.291	0.203	0.147	0.367	–	0.304	5-fold CV

All values were extracted from their papers, except in some cases where results were not present in the own paper: [(1)] TieNet [144] results were presented in Liu et al. [92] as a baseline; [(2)] Xue et al. 2018 [154] results in the findings section were presented in Xue et al. 2019 [153] as a baseline. [(3)] CLARA [16] results are from the fully automatic version.

imaging experts, for example, starting with chest X-rays, which is the medical image type with more datasets available and research done. A standard protocol should facilitate fair comparisons between studies and allow to assess how close a model is to meet standard criteria for deployment in a clinical setting.

The expertise of the human evaluators is an important factor to consider as well. It stands to reason that the judgment of a board-certified radiologist with years of experience should be more reliable than the judgment of a physician with limited experience. Similarly, the consensus of a team of radiologists should be preferred over a single radiologist. In the same line, measuring the inter-agreement of several radiologists can help to better assess the difficulty of the task itself [65]. If radiologists tend to disagree more, this may indicate an inherent ambiguity in the task that could be the explanation for the possible underperformance of a given model.

Automatic metrics for medical correctness. Having a proper expert evaluation is desirable. However, it is not feasible to ask radiologists to manually evaluate hundreds of machine-generated reports every time a small tweak in a model is performed. Instead, one would like to have one or more automatic metrics positively correlated with expert human evaluation, in order to speed up the model design and testing cycle. We found that more than 70% of the works reviewed (29 out of 40) limit the automatic report evaluation to traditional NLP metrics such as BLEU, ROUGE-L, or CIDEr, which are not designed to evaluate a report from a medical correctness

point of view [11, 17, 92, 107, 108, 161]. Furthermore, these evaluation methods have been recently contested in other NLP tasks [96, 111, 137, 138]. Some works tried to remedy this limitation by devising their own auxiliary metrics to evaluate medical correctness to some degree [7, 16, 61, 67, 85, 92, 98, 148, 150, 154, 162], which are interesting approaches. We highlight the metric MIRQI proposed very recently by Zhang et al. [162], which is very similar to SPICE [9] as described in Section 5.4.2, as it attempts to build a graph capturing abnormalities and their relations and attributes. We believe this is the most sophisticated metric for medical correctness found in the literature, and great ideas can be adopted from it.

Unfortunately, all the proposed metrics lack validation by medical experts, as none of the papers presents the results of a study assessing the correlation between the proposed metric and expert medical judgment. Thus, finding one or more golden automatic metrics for medical correctness remains an open problem. To solve it, the precision and accuracy of a report are critical in the medical domain and need to be captured [11, 107, 161], whereas other aspects such as natural language fluency should probably weigh less in importance. We believe designing and validating such metrics is a clear avenue for future research, with the potential to have a significant impact on the field.

Improve explainability. To build trust in an AI system, a desirable feature is the ability to provide clear and coherent explanations for its decisions [2, 34]. This is particularly relevant in the healthcare domain, where decisions have to be made with extreme caution since the patient's health is at stake. Thus, high levels of transparency, interpretability, and accountability are required to justify the outputs delivered, align to the expert's expectations, and acquire their trust [113, 133, 134].

Only two papers reviewed [38, 126] have explainability as a primary focus, as discussed in Section 5.3, though one of them [38] does not provide an explanation for the report. Additionally, some works mention some form of local explainability in their models, but always as a secondary output and giving it a rather superficial treatment, with no rigorous evaluation. In the absence of empirical results across all papers, we cannot draw conclusions about which explanation techniques are better or worse. Thus, a potential avenue for future research is explainability with a more rigorous and empirical focus, and possibly including other approaches, such as global explanations, uncertainty, or more, which may be necessary for clinicians [134]. We believe this research avenue will benefit from the feedback and evaluation of medical imaging experts, who are the end-users of these systems. What would be a suitable explanation for a radiologist? In a multi-sentence report, how should the explanation be structured? An expert's opinion is valuable for answering these and other questions, and ultimately for assessing the explanation.

New learning strategies and architectures. If the ultimate goal is to have a model that learns to generate accurate and useful medical reports, the optimization strategy employed should be designed to guide the model in this direction. As we saw in Section 5.2.6, most papers used teacher-forcing, a training strategy that is domain agnostic and thus suboptimal for the medical domain [161]. Similarly, a few papers used reinforcement learning [67, 85, 87, 92, 151] with traditional NLP metrics as rewards, which are not designed for medicine either. Only Liu et al. [92] included a domain-specific reward that explicitly promotes medical correctness. Unfortunately, a manual inspection of several generated reports conducted by the authors revealed that the model was missing positive findings (low recall) as well as failing to provide accurate descriptions of the positive findings detected.

Given these reasons, there is still room for finding better optimization strategies for image-based medical report generation. In this regard, reinforcement learning appears to be the most promising training paradigm to explore, as illustrated by the work of Zhang et al. [161] on factual correctness optimization in a related medical task. If a robust medical correctness metric is developed (as

previously discussed in this section), then the metric could be used as a reward in a reinforcement learning setting to teach the model to generate reports that are medically correct.

Other image modalities and body regions are less explored. Most research has concentrated on chest X-rays, as 24 out of 40 papers focus their study on this image type. This modality presents a very specific nature and different characteristics from other imaging studies. For example, when a radiologist reads a chest X-ray, the focus is on the underlying anatomy and identification of possible areas of distortion based on different densities of the image. On the other hand, when analyzing a PET image, the focus is on detecting areas of increased radiotracer activity; for MRI scans the radiologist may review several images obtained with different configurations at the same time; and for each other modality there may be more specific conditions. Hence, the results shown here are highly biased toward chest X-rays, which will not necessarily extrapolate to other scenarios.

Notice there are datasets with multiple image types or body parts, namely ImageCLEF caption [35, 39], ROCO [105], and PEIR Gross [68], as it was mentioned in Section 5.1. However, we believe their broad nature, i.e., the inclusion of many types and regions simultaneously, may be a drawback when trying to apply an advanced deep learning approach, for three main reasons. First, it is more difficult to include specific domain knowledge in the models, as the knowledge should cover all modalities and body parts. Second, assessing medical correctness is more complicated, since domain knowledge is needed to design these metrics, as noted in Section 5.4. Third, it would be more challenging to provide interpretability for the model, as the explanations should cover all modalities. Ultimately, we believe better solutions can be achieved by designing them for a specific problem and setting. In conclusion, there is a clear opportunity to extend research into other image types and body regions by raising new collections with other image types, evaluating the same methods in different modalities, or further covering the existing datasets.

Explore more alternatives to include domain knowledge. As we saw in Section 5.2.4, the approaches explored in the literature for incorporating domain knowledge into models are (1) the use of graph neural networks at the visual component level and (2) the use template databases curated with expert knowledge—in addition to the widespread use of auxiliary tasks, which can be viewed as a way of domain knowledge transfer as well. However, other approaches remain unexplored. A recent survey by Xie et al. [149] synthesizing over 270 papers on domain knowledge for deep-learning-based medical image analysis presents interesting ideas that could be applicable to the report generation setting. For example, curriculum learning [15] and self-paced learning [80] could be used to imitate the learning curve from easier to harder instances that radiologists go through when they learn to interpret and diagnose images. Also, the use of handcrafted algorithms to extract visual features that better capture what radiologists focus on in an image could be used, which many works have verified to have synergistic effects in combination with the features learned by the CNN [149]. This would improve the quality of the visual component and potentially translate into better reports. Studying how imaging experts analyze an image, how they focus the attention to different regions of the image as needed, could be useful to inspire innovations in model architectures in order to emulate that process.

Medical human-AI interaction. Most reviewed works leave aside important aspects pertaining the model's integration in a real clinical setting and its interaction with clinicians as an AI assistant. Besides high levels of accuracy, there are other needs a system should aim to meet in a medical human-AI collaboration workflow. For example, Cai et al. [20] argue that clinicians should have transparent information about the model's overall strengths and weaknesses, its subjective point of view, its overall design objective, and how exactly it uses the information to derive a final diagnosis. Also, Amershi et al. [8] proposed and validated several design guidelines for general human-AI

interaction that can be relevant in the context of automatic report generation, such as *Make clear why the system did what it did* via explanations, and *Support efficient correction* by making it easy to edit, refine, or recover when the AI system is wrong. Among all papers reviewed, only CLARA [16] targets an explicit workflow with human interaction, in which a report is generated cooperatively by a human who types some preliminary text and the system autocompletes the rest.

Also, there are potential use cases that an AI assistant for report generation can face in routine practice that are not addressed in the reviewed literature. For example, (1) *open-ended* ***visual question answering (VQA)***: instead of a full report with too many details, a clinician might be interested in the model's opinion on a specific aspect of the image(s). This query could be expressed as a natural language question that the model would have to answer, which would require a model with open-ended VQA capabilities. Although this is a different task than report generation, we believe the latter could be approached as giving answers to a sequence of questions from physicians, allowing a richer interaction between the expert and the system. The multiple ImageCLEF challenges involving a medical VQA task [13, 14, 50] and the recently published PathVQA dataset [53] could be helpful in exploring this direction. (2) *Reporting temporal information*: sometimes clinicians are interested in the evolution of a health condition by analyzing a sequence of imaging snapshots over time, rather than describing a single image. None of the surveyed papers considers this use case. (3) *Quantitative radiology*: in some cases a clinician might be interested in specific numerical measurements to further assess the patient's condition, for example, the degree of a certain property in the tissues [64]. This adds more complexity to the problem, since models would need the ability to make these accurate numerical measurements, in addition to interpreting them through words in the generated report. In sum, there may be different ways to fulfill the report generation task, and we believe researchers should aim to find the most useful approaches for clinicians in each specific environment.

7 LIMITATIONS

The main limitations of this survey are two. First, new papers on report generation from medical images are published relatively often; we tried to be as comprehensive as possible and include all of them, but we do not rule out that some papers may have been missed. Second, we left out of the analysis works from related tasks, such as disease classification, report summarizing, or medical image segmentation. These topics may have interesting approaches or insights on how to improve the visual features generated, how to optimize the text generation, evaluation techniques, and more.

8 CONCLUSIONS

In this work, we have reviewed the state of research in deep-learning-based methods for automatic report generation from medical images, in terms of different key aspects. First, we described the report and classification **datasets** available and commonly used in the literature, totaling 27 collections, which cover different image modalities and body parts, and include useful tags and localization information. Second, we presented an analysis of **model designs** in terms of standard practices, inputs and outputs, visual components, language components, domain knowledge, auxiliary tasks, and optimization strategies. We cannot recommend an optimal model design due to the lack of proper evaluations, but several guidelines can be inferred. For instance, a robust visual component should make use of CNNs and would certainly benefit from training in auxiliary medical image tasks. Also, complementing the visual input with semantic information via tags or input text (e.g., the report's *indication* section) or access to a template database generally improves the language component's performance. Multitask learning to integrate the supervision of multiple tasks and reinforcement learning to directly optimize for factual correctness or other metrics of interest

in generated reports appear as the most promising optimization approaches. Third, we analyzed the **interpretability** approaches employed in the literature and found that many models provide a secondary output that can be used as a local explanation, either by providing a feature importance map, by providing a counter-factual example, or by increasing the system's transparency. However, only two works focused explicitly on studying this concern, by discussing extensively and providing formal evaluations. Additionally, many other approaches can be explored, and hence this remains a heavily understudied aspect of this task. Fourth, we discussed usual practices regarding **evaluation metrics**, and we found that most models are only evaluated with traditional n-gram-based NLP metrics not designed for medicine, which are not able to capture the essential medical facts in a written report. Next, we presented a comparison of papers' **performance results** on IU X-Ray, the most frequently used dataset, but limited to said NLP metrics that papers report, making us unable to judge models from a medical perspective.

Lastly, we identified **challenges** in the field that none of the reviewed papers has successfully addressed, and we proposed avenues for future research where we believe possible solutions could be found. The main challenges lay in improving the evaluation methods employed, by developing a *standard protocol for expert evaluation* and *automatic metrics for medical correctness*. Other important aspects are improving the *explainability of models* and considering the *medical human-AI interaction*. We intend this survey to serve as an entry point for researchers who want an overview of the current advances in the field and also to raise awareness of critical problems that future research should focus on, with the end goal of developing mature and robust technologies that can bring value to healthcare professionals and patients in real clinical settings.

A APPENDIX

A.1 Datasets

Next, we include Table 8 with the main highlights of all datasets, including both public and proprietary, and Table 9 with details of the additional information provided by each collection.

Table 8. Datasets Used in the Literature

Dataset	Year	Image Type	# Images	# Reports	# Patients	Used by Papers
Public report datasets						
IU X-ray [28]	2015	Chest X-ray	7,470	3,955	3,955	[16, 36, 40, 48, 61, 67, 68, 85, 86, 89, 92, 120, 123, 132, 144, 150, 151, 153–156, 162]
MIMIC-CXR [69, 70]	2019	Chest X-ray	377,110	227,827	65,379	[92]
PadChest(sp) [19]	2019	Chest X-ray	160,868	109,931	67,625	None(5)
ImageCLEF Caption 2017 [35]	2017	Biomedical(2)	184,614	184,614	–	[51]
ImageCLEF Caption 2018 [39]	2018	Biomedical(2)	232,305	232,305	–	None(5)
ROCO [105]	2018	Multiple radiology(3)	81,825	81,825	–	None(5)
PEIR Gross [68]	2017	Gross lesions	7,442	7,442	–	[68]
INBreast(pt) [99]	2012	Mammography X-ray	410	115	115	[87, 128]
STARE [58]	1975	Retinal fundus	400	400	–	None(5)
RDIF(1) [95]	2019	Kidney biopsy	1,152	144	144	[95]
Private Report Datasets						
CX-CHR(ch) [67, 85, 86]	2018	Chest X-ray	45,598	35,609	35,609	[67, 85, 86]
TJU(ch) [44]	2019	Chest X-ray	19,985	19,985	–	[44]
Hip fracture [37, 38]	2017	Hip X-ray	53,279	4,010	26,639	[38]
Ultrasound [157, 158]	2018	Gallbladder, kidney, and liver ultrasound(4)	4,302	4,302	–	[157, 158]
Fetal Ultrasound [7]	2019	Fetal ultrasound(4)	2,800	2,800	–	[7]
CINDRAL [94]	2018	Cervical neoplasm WSI	1,000	1,000	50	[94]
BCIDR [163]	2017	Bladder biopsy	1,000	1,000	32	[163]
Continuous wave [98]	2016	Continuous wave Doppler echocardiography	722	10,479	–	[98]
Public Classification Datasets						
CheXpert [63]	2019	Chest X-ray	224,316	0	65,240	[156, 162]
ChestX-ray14 [143]	2017	Chest X-ray	112,120	0	30,805	[16, 67, 86, 89, 144, 151, 153]
LiTS [25]	2017	Liver CT scans	200	0	–	[131]
ACM Biomedia 2019 [55]	2019	Gastrointestinal tract (4)	14,033	0	–	[49]
DIARETDB0 [73]	2006	Retinal fundus	130	0	–	[148]
DIARETDB1 [72]	2007	Retinal fundus	89	0	–	[148]
Messidor [1, 27]	2013	Retinal fundus	1,748	0	874	[148]
DDSM [54]	2001	Mammography X-ray	10,480	0	–	[76]
Private Classification Datasets						
MRI Spine [47]	2018	Spine MRI scans	≥253	0	253	[47]

WSI stands for Whole Slide Images. All reports are written in English, except those marked with (sp) which are in Spanish, with (ch) in Chinese, and (pt) in Portuguese. Other notes, (1): the RDIF dataset is pending release. (2): for the ImageCLEF datasets, images were extracted from PubMed Central papers and filtered with an automatically to keep only clinical images, then it contains samples from other domains. (3): contains multiple modalities, namely CT, Ultrasound, X-Ray, Fluoroscopy, PET, Mammography, MRI, Angiography and PET-CT. (4): the images are frames extracted from videos. (5): none of the papers reviewed used this dataset.

Table 9. Additional Data Contained in Each Dataset

Dataset	Text	Tags	Tags Annotation Method	Localization
Public Report Datasets				
IU X-ray [28]	Indication	(1) MeSH and RadLex concepts (2) MeSH concepts	(1) Manual (2) MTI and MetaMap	–
MIMIC-CXR [69, 70]	Comparis, Indicat	14 CheXpert labels	CheXpert labeler and NegBio	–
PadChest [19]	–	297 labels (findings, diagnoses, and anatomic)	27% manual, rest by RNN	–
ImageCLEF Caption 2017 [35]	–	UMLS tags	Quick-UMLS	–
ImageCLEF Caption 2018 [39]	–	UMLS tags	Quick-UMLS	–
ROCO [105]	–	UMLS tags	Quick-UMLS	–
PEIR Gross [68]	–	Top words	Top TF-IDF scores	–
INBreast [99]	–	Abnormalities	Manual	Abnormality contours
STARE [58]	–	Levels for 39 conditions and presence of 13 diagnostics	Manual	–
RDIF [95]	Indication	–	–	–
Private Report Datasets				
CX-CHR [67, 85, 86]	–	–	–	–
TJU [44]	–	Top abnormality words	40 most frequent words	–
Hip fracture [37, 38]	–	(1) Fracture presence, (2) fracture location and character	(1) CNN [37], (2) manual [38]	–
Ultrasound [157, 158]	–	Organ and disease	Manual	Organ bounding boxes
Fetal Ultrasound [7]	–	Body part	Manual	–
CINDRAL [94]	–	Severity level for 4 attributes and diagnosis label	Manual	–
BCIDR [163]	–	Disease status (4 possible)	Manual	–
Continuous wave [98]	–	Valve types	Manual	–
Public Classification Datasets				
CheXpert [63]	–	14 CheXpert labels	CheXpert labeler	–
ChestX-ray14 [143]	–	14 disease labels	DNorm and MetaMap	Disease bounding boxes for 880 images
LiTS [25]	–	–	–	Liver and tumor segmentation masks
Gastrointestinal challenge [55]	–	16 labels (e.g., anatomic, pathological or surgery findings)	Manual	–
DIARETDB0 [73]	–	DR severity level	Manual	Abnormality contours
DIARETDB1 [72]	–	DR severity level	Manual	Abnormality contours
Messidor [1, 27]	–	DR severity level	Manual	–
DDSM [54]	–	Density level	Manual	Abnormalities at pixel level
Private Classification Datasets				
MRI Spine [47]	–	–	–	Diseases and body parts at pixel level

MeSH [115] and RadLex [81] are sets of medical concepts. MTI [100], MetaMap [10], CheXpert labeler [63], NegBio [106], Quick-UMLS [125] and DNorm [84] are automatic labeler tools. *Manual* means manually annotated by experts. In all cases, the localization information was manually annotated by experts.

A.2 Auxiliary Tasks

Table 10 presents the categories of auxiliary tasks identified in the literature and which papers implemented them.

Table 10. Summary of Auxiliary Tasks Used in the Literature

Auxiliary Task	Used by Papers
Multi-label classification	[16, 44, 48, 49, 67, 68, 86, 89, 120, 128, 132, 144, 151, 155, 156, 162]
Single-label classification	[7, 38, 40, 51, 76, 94, 98, 126, 157, 158, 163]
Sentence classification (normal/abnormal/stop)	[48, 67, 150]
Segmentation	[47, 131]
Object detection	[76, 157]
Attention weights regularization	[95, 155]
Embedding-to-embedding matching	[98, 155]
Doc2vec	[98]
Text autoencoder	[126, 132]
GAN cycle-consistency	[126]

A.3 Optimization Strategies

Table 11 presents the categories of optimization strategies identified in the literature and which papers implemented them.

Table 11. Summary of Optimization Strategies Used in the Literature

Category	Optimization Strategy	Used by Papers
Visual Component	Pretrain in ImageNet	[7, 40, 51, 61, 67, 85, 89, 94, 95, 98, 123, 144, 153–155, 157, 158]
	Train in auxiliary medical image tasks	[7, 16, 38, 44, 47–49, 51, 67, 76, 86, 89, 94, 120, 128, 131, 132, 151, 155–158, 162, 163]
	Train in report generation (end-to-end)	[40, 48, 51, 68, 87, 92, 131, 132, 144, 153, 155, 163]
Report Generation	Teacher-forcing	[7, 16, 36, 38, 40, 44, 48, 51, 61, 67, 68, 86, 87, 89, 95, 120, 123, 126, 128, 131, 132, 144, 148, 150, 153–158, 162, 163]
	Reinforcement learning	[67, 85, 87, 92, 151]
Other Losses or Training Strategies	Multi-task learning	[48, 67, 68, 76, 86, 94, 95, 126, 131, 132, 144, 155, 157, 163]
	Attention weights regularization	[95, 155]
	Contrastive loss	[155]
	Regression loss	[76, 98, 157]
	Autoencoder	[126, 132]
	GAN	[47, 87, 126]

REFERENCES

[1] Michael D. Abràmoff, James C. Folk, Dennis P. Han, Jonathan D. Walker, David F. Williams, Stephen R. Russell, Pascale Massin, Beatrice Cochener, Philippe Gain, Li Tang, Mathieu Lamard, Daniela C. Moga, Gwénolé Quellec, and Meindert Niemeijer. 2013. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmology* 131, 3 (2013), 351–357.

[2] A. Adadi and M. Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 9505–9515.

[4] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable machine learning in healthcare. In *Proc. of the 2018 ACM Intl. Conf. on Bioinformatics, Computational Biology, and Health Informatics (BCB'18)*. ACM, New York, NY, 559–560.

[5] Kentaro Akazawa, Ryo Sakamoto, Satoshi Nakajima, Dan Wu, Yue Li, Kenichi Oishi, Andreia V. Faria, Kei Yamada, Kaori Togashi, Constantine G. Lyketsos, Michael I. Miller, and Susumu Mori. 2019. Automated generation of radiologic descriptions on brain volume changes from T1-weighted MR images: Initial assessment of feasibility. *Frontiers in Neurology* 10 (2019), 7.

[6] Imane Allaouzi, M. Ben Ahmed, B. Benamrou, and M. Ouardouz. 2018. Automatic caption generation for medical images. In *Proc. of the 3rd Intl. Conf. on Smart City Applications (SCA'18)*. ACM, New York, NY, Article 86, 6 pages.

[7] Mohammad Alsharid, Harshita Sharma, Lior Drukker, Pierre Chatelain, Aris T. Papageorghiou, and J. Alison Noble. 2019. Captioning ultrasound images automatically. In *Medical Image Computing and Computer Assisted Intervention (MICCAI'19)*. Springer Intl. Publishing, Cham, 338–346.

[8] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-AI interaction. In *Proc. of the 2019 CHI Conf. on Human Factors in Computing Systems (CHI'19)*. ACM, 1–13.

[9] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. In *Computer Vision (ECCV'16)*. Springer Intl. Publishing, Cham, 382–398.

[10] Alan R. Aronson and François-Michel Lang. 2010. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17, 3 (2010), 229–236.

[11] Zaheer Babar, Twan van Laarhoven, Fabio Massimo Zanzotto, and Elena Marchiori. 2021. Evaluating diagnostic content of AI-generated radiology reports of chest X-rays. *Artificial Intelligence in Medicine* 116 (2021), 102075. https://doi.org/10.1016/j.artmed.2021.102075

[12] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. ACL, 65–72.

[13] Asma Ben Abacha, Vivek V. Datla, Sadid A. Hasan, Dina Demner-Fushman, and Henning Müller. 2020. Overview of the VQA-med task at ImageCLEF 2020: Visual question answering and generation in the medical domain. In *CLEF 2020 Working Notes (CEUR Workshop Proceedings)*.

[14] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. In *CLEF2019 Working Notes (CEUR Workshop Proceedings)*.

[15] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proc. of the 26th Annual Intl. Conf. on Machine Learning (ICML'09)*. ACM, 41–48.

[16] Siddharth Biswal, Cao Xiao, Lucas M. Glass, Brandon Westover, and Jimeng Sun. 2020. CLARA: Clinical report autocompletion. In *Proc. of the Web Conf. 2020 (WWW'20)*. ACM, New York, NY, 541–550.

[17] William Boag, Tzu-Ming Harry Hsu, Matthew Mcdermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. 2020. Baselines for chest X-ray report generation. In *Proc. of the Machine Learning for Health NeurIPS Workshop (Proc. of Machine Learning Research)*, Vol. 116. PMLR, 126–140.

[18] Milosavljević Branko, Boberić Danijela, and Surla Dušan. 2010. Retrieval of bibliographic records using Apache Lucene. *Electronic Library* 28, 4 (Jan. 2010), 525–539.

[19] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2019. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv:1901.07441* (2019).

[20] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proc. of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[21] Rich Caruana. 1997. Multitask learning. *Machine Learning* 28, 1 (1997), 41–75.

[22] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[23] Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP'14)*. ACL, 740–750.

[24] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 539–546.

[25] P. Christ, F. Ettlinger, F. Grün, J. Lipkova, and G. Kaissis. 2017. Lits-liver tumor segmentation challenge. *ISBI and MICCAI* (2017).

[26] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*.

[27] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton, and Klein Jc. 2014. Feedback on a publicly distributed image database: The Messidor database. *Image Analysis & Stereology* 33, 3 (2014), 231–234.

[28] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* 23, 2 (2015), 304–310.

[29] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conf. on Computer Vision and Pattern Recognition*. 248–255.

[30] Michael Denkowski and Alon Lavie. 2010. Extending the meteor machine translation evaluation metric to the phrase level. In *Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the ACL (HLT'10)*. ACL, 250–253.

[31] Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proc. of the 6th Workshop on Statistical Machine Translation (WMT'11)*. ACL, 85–91.

[32] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proc. of the 9th Workshop on Statistical Machine Translation*. ACL, 376–380.

[33] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *stat* 1050 (2017), 2.

[34] F. K. Došilović, M. Brčić, and N. Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st Intl. Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO'18)*. 0210–0215.

[35] Carsten Eickhoff, Immanuel Schwall, Alba García Seco de Herrera, and Henning Müller. 2017. Overview of ImageCLEFcaption 2017 - The image caption prediction and concept extraction tasks to understand biomedical images. In *CLEF2017 Working Notes (CEUR Workshop Proceedings)*.

[36] Gaurav O. Gajbhiye, Abhijeet V. Nandedkar, and Ibrahima Faye. 2020. Automatic report generation for chest X-ray images: A multilevel multi-attention approach. In *Computer Vision and Image Processing*. Springer, Singapore, 174–182.

[37] William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P. Bradley, and Lyle J. Palmer. 2017. Detecting hip fractures with radiologist-level performance using deep neural networks. *arXiv:1711.06504* (2017).

[38] W. Gale, L. Oakden-Rayner, G. Carneiro, L. J. Palmer, and A. P. Bradley. 2019. Producing radiologist-quality reports for interpretable deep learning. In *2019 IEEE 16th Intl. Symposium on Biomedical Imaging (ISBI'19)*. 1275–1279.

[39] Alba García Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, and Henning Müller. 2018. Overview of the ImageCLEF 2018 caption prediction tasks. In *CLEF2018 Working Notes (CEUR Workshop Proceedings)*.

[40] Aydan Gasimova. 2019. Automated enriched medical concept generation for chest X-ray images. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer Intl. Publishing, Cham, 83–92.

[41] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2672–2680.

[42] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5–6 (2005), 602–610.

[43] Mara Graziani, Vincent Andrearczyk, and Henning Müller. 2018. Regression concept vectors for bidirectional explanations in histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer Intl. Publishing, Cham, 124–132.

[44] M. Gu, X. Huang, and Y. Fang. 2019. Automatic generation of pulmonary radiology reports with semantic tags. In *2019 IEEE 11th Intl. Conf. on Advanced Infocomm Technology (ICAIT'19)*. 162–167.

[45] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–42.

[46] David Gunning. 2017. Explainable artificial intelligence (XAI). DOI: https://doi.org/10.1609/aimag.v40i2.2850

[47] Zhongyi Han, Benzheng Wei, Stephanie Leung, Jonathan Chung, and Shuo Li. 2018. Towards automatic report generation in spine radiology using weakly supervised framework. In *Medical Image Computing and Computer Assisted Intervention (MICCAI'18)*. Springer Intl. Publishing, Cham, 185–193.

[48] Philipp Harzig, Yan-Ying Chen, Francine Chen, and Rainer Lienhart. 2019. Addressing data bias problems for chest X-ray image report generation. arXiv abs/1908.02123 (2019).

[49] Philipp Harzig, Moritz Einfalt, and Rainer Lienhart. 2019. Automatic disease detection and report generation for gastrointestinal tract examination. In *Proc. of the 27th ACM Intl. Conf. on Multimedia (MM'19)*. ACM, New York, NY, 2573–2577.

[50] Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew Lungren, and Henning Müller. 2018. Overview of the ImageCLEF 2018 medical domain visual question answering task. In *CLEF2018 Working Notes (CEUR Workshop Proceedings)*.

[51] Sadid A. Hasan, Yuan Ling, Joey Liu, Rithesh Sreenivasan, Shreya Anand, Tilak Raj Arora, Vivek Datla, Kathy Lee, Ashequl Qadir, Christine Swisher, and Oladimeji Farri. 2018. Attention-based medical caption generation with image modality classification and clinical concept mapping. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer Intl. Publishing, Cham, 224–230.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'16)*. 770–778.

[53] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. PathVQA: 30000+ questions for medical visual question answering. *arXiv:2003.10286* (2020).

[54] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and P. Kegelmeyer. 2001. The digital database for screening mammography. In *Proc of the 5th Intl. Workshop on Digital Mammography*, Vol. 58, M. J. Yaffe, ed. Medical Physics Publishing, 212–218.

[55] Steven Hicks, Michael Riegler, Pia Smedsrud, Trine B. Haugen, Kristin Ranheim Randel, Konstantin Pogorelov, Håkon Kvale Stensland, Duc-Tien Dang-Nguyen, Mathias Lux, Andreas Petlund, Thomas de Lange, Peter Thelin Schmidt, and Pål Halvorsen. 2019. ACM multimedia BioMedia 2019 grand challenge overview. In *Proc. of the 27th ACM Intl. Conf. on Multimedia (MM'19)*. ACM, 2563–2567.

[56] Steven Alexander Hicks, Konstantin Pogorelov, Thomas de Lange, Mathias Lux, Mattis Jeppsson, Kristin Ranheim Randel, Sigrun Eskeland, Pål Halvorsen, and Michael Riegler. 2018. Comprehensible reasoning and automated reporting of medical examinations based on deep learning analysis. In *Proc of the 9th ACM Multimedia Systems Conference (MMSys'18)*. ACM, 490–493.

[57] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[58] A. Hoover. 1975. STARE database. http://www.ces.clemson.edu/~ahoover/stare.

[59] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861* (2017).

[60] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'17)*. 4700–4708.

[61] Xin Huang, Fengqi Yan, Wei Xu, and Maozhen Li. 2019. Multi-attention and incorporating background information model for chest X-ray image report generation. *IEEE Access* 7 (2019), 154808–154817.

[62] Eui Jin Hwang, Sunggyun Park, Kwang-Nam Jin, Jung Im Kim, So Young Choi, Jong Hyuk Lee, Jin Mo Goo, Jaehong Aum, Jae-Joon Yim, Julien G. Cohen, Gilbert R. Ferretti, Chang Min Park, for the DLAD Development, and Evaluation Group. 2019. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Network Open* 2, 3 (03 2019), e191095–e191095.

[63] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proc. of the AAAI Conf. on Artificial Intelligence*, Vol. 33. Association for the Advancement of Artificial Intelligence (AAAI), 590–597.

[64] E. F. Jackson. 2018. Quantitative Imaging: The translation from research tool to clinical practice. *Radiology* 286, 2 (2018), 499.

[65] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis Langlotz, and Pranav Rajpurkar. 2021. RadGraph: Extracting clinical entities and relations from radiology reports. In *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. https://openreview.net/forum?id=pMWtc5NKd7V.

[66] Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proc. of the 2019 Conf. of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. ACL.

[67] Baoyu Jing, Zeya Wang, and Eric Xing. 2019. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In *Proc of the 57th Annual Meeting of the ACL*. ACL, 6570–6580.

[68] Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Proc. of the 56th Annual Meeting of the ACL (Volume 1: Long Papers)*. ACL, 2577–2586.

[69] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv:1901.07042* (2019).

[70] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* 6, 1 (Dec. 2019), 317.

[71] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. 1996. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* 4 (1996), 237–285.
[72] R. V. J. P. H. Kälviäinen and H. Uusitalo. 2007. DIARETDB1 diabetic retinopathy database and evaluation protocol. In *Medical Image Understanding and Analysis*, Vol. 2007. Citeseer, 61.
[73] Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iiris Sorri, Hannu Uusitalo, Heikki Kälviäinen, and Juhani Pietilä. 2006. DIARETDB0: Evaluation database and methodology for diabetic retinopathy algorithms. *Machine Vision and Pattern Recognition Research Group* 73 (2006), 1–17.
[74] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. 2020. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review* (April 2020), 1–62.
[75] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proc. of Machine Learning Research*, Vol. 80. PMLR, 2668–2677.
[76] Pavel Kisilev, Eli Sason, Ella Barkan, and Sharbell Hashoul. 2016. Medical image description using multi-task-loss CNN. In *Deep Learning and Data Labeling for Medical Applications*. Springer Intl. Publishing, Cham, 121–129.
[77] Nikos Komodakis and Sergey Zagoruyko. 2017. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*.
[78] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. 2019. Do better imagenet models transfer better? In *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR'19)*. IEEE Computer Society, Los Alamitos, CA, 2656–2666.
[79] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 1097–1105.
[80] M. P. Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc., 1189–1197.
[81] Curtis P. Langlotz. 2006. RadLex: A new method for indexing online educational materials. *Radiographics: A Review Publication of the Radiological Society of North America, Inc.* 26, 6 (2006), 1595.
[82] Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proc of the 2nd Workshop on Statistical Machine Translation (StatMT'07)*. ACL, 228–231.
[83] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Intl. Conf. on Machine Learning*. 1188–1196.
[84] Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics* 57 (2015), 28–37.
[85] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Proc. of the 32nd Intl. Conf. on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, 1537–1547.
[86] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proc. of the AAAI Conf. on Artificial Intelligence*, Vol. 33. 6666–6673.
[87] Jiyun Li and Yongliang Hong. 2019. Label generation system based on generative adversarial network for medical image. In *Proc. of the 2nd Intl. Conf. on Artificial Intelligence and Pattern Recognition (AIPR'19)*. ACM, 78–82.
[88] Jiwei Li, Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proc of the 53rd Annual Meeting of the ACL and the 7th Intl. Joint Conf. on Natural Language Processing (Volume 1: Long Papers)*. ACL, 1106–1115.
[89] Xin Li, Rui Cao, and Dongxiao Zhu. 2019. Vispi: Automatic visual perception and interpretation of chest X-rays. *arXiv:1906.05190* (2019).
[90] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. ACL, 74–81.
[91] Zachary C. Lipton. 2018. The mythos of model interpretability. *Communications of the ACM* 61, 10 (2018), 36–43.
[92] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest X-ray report generation. In *Machine Learning for Healthcare Conference (Proc of Machine Learning Research)*, Vol. 106. PMLR, 249–269.
[93] Samira Loveymi, Mir Hossein Dezfoulian, and Muharram Mansoorizadeh. 2020. Generate structured radiology report from CT images using image annotation techniques: Preliminary results with liver CT. *Journal of Digital Imaging* 33, 2 (April 2020), 375–390.
[94] Kai Ma, Kaijie Wu, Hao Cheng, Chaochen Gu, Rui Xu, and Xinping Guan. 2018. A pathology image diagnosis network with visual interpretability and structured diagnostic report. In *Neural Information Processing*. Springer Intl. Publishing, Cham, 282–293.
[95] Sam Maksoud, Arnold Wiliem, Kun Zhao, Teng Zhang, Lin Wu, and Brian Lovell. 2019. CORAL8: Concurrent object regression for area localization in medical image panels. In *Medical Image Computing and Computer Assisted Intervention (MICCAI'19)*. Springer Intl. Publishing, Cham, 432–441.

[96] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4984–4997. https://doi.org/10.18653/v1/2020.acl-main.448

[97] Maram Mahmoud A. Monshi, Josiah Poon, and Vera Chung. 2020. Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine* 106 (2020), 101878.

[98] Mehdi Moradi, Yufan Guo, Yaniv Gur, Mohammadreza Negahdar, and Tanveer Syeda-Mahmood. 2016. A cross-modality neural network transform for semi-automatic medical image annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI'16)*. Springer Intl. Publishing, Cham, 300–307.

[99] Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S. Cardoso. 2012. Inbreast: Toward a full-field digital mammographic database. *Academic Radiology* 19, 2 (2012), 236–248.

[100] J. G. Mork, A. J. J. Yepes, and A. R. Aronson. 2013. The NLM medical text indexer system for indexing biomedical literature. In *CEUR Workshop Proceedings*, Vol. 1094.

[101] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* 32, 2 (2020), 604–624. https://europepmc.org/article/med/32324570.

[102] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the ACL*. ACL, 311–318.

[103] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proc. of the 30th Intl. Conf. on Intl. Conf. on Machine Learning - Volume 28 (ICML'13)*. JMLR.org, III–1310–III–1318.

[104] John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. 2019. A survey on biomedical image captioning. In *Proc. of the 2nd Workshop on Shortcomings in Vision and Language*. ACL, 26–36.

[105] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. 2018. Radiology objects in COntext (ROCO): A multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer Intl. Publishing, Cham, 180–189.

[106] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. Negbio: A high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings* 2018 (2018), 188.

[107] Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. 2021. Clinically correct report generation from chest X-rays using templates. In *Machine Learning in Medical Imaging*, Chunfeng Lian, Xiaohuan Cao, Islem Rekik, Xuanang Xu, and Pingkun Yan (Eds.). Springer International Publishing, Cham, 654–663. https://doi.org/10.1007/978-3-030-87589-3_67

[108] Pablo Pino, Denis Parra, Pablo Messina, Cecilia Besa, and Sergio Uribe. 2020. Inspecting state of the art performance and NLP metrics in image-based medical report generation. arXiv preprint arXiv:2011.09257 (2020). In LXAI at NeurIPS 2020.

[109] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. 2019. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 3347–3357.

[110] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv:1711.05225* (2017).

[111] Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics* 44, 3 (2018), 393–401. https://doi.org/10.1162/coli_a_00322

[112] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 91–99.

[113] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest. 2020. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology: Artificial Intelligence* 2, 3 (2020), e190043.

[114] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proc. of the 22nd ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD'16)*. ACM, 1135–1144.

[115] Frank B. Rogers. 1963. Medical subject headings. *Bulletin of the Medical Library Association* 51, 1 (1963), 114–116.

[116] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI'15)*. Springer Intl. Publishing, Cham, 234–241.

[117] David A. Rosman, Judith Bamporiki, Rebecca Stein-Wexler, and Robert D. Harris. 2019. Developing diagnostic radiology training in low resource countries. *Current Radiology Reports* 7, 9 (2019), 27.

[118] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV'17)*. 618–626.

[119] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics* 22, 5 (2017), 1589–1604.

[120] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M. Summers. 2016. Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'16)*. 2497–2506.

[121] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proc. of the 34th Intl. Conf. on Machine Learning - Volume 70 (ICML'17)*. JMLR.org, 3145–3153.

[122] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556* (2014).

[123] Sonit Singh, Sarvnaz Karimi, Kevin Ho-Shon, and Len Hamey. 2019. From chest X-rays to radiology reports: A multimodal machine learning approach. In *2019 Digital Image Computing: Techniques and Applications (DICTA'19)*. IEEE, 1–8.

[124] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: Removing noise by adding noise. *arXiv:1706.03825* (2017).

[125] Luca Soldaini and Nazli Goharian. 2016. Quickumls: A fast, unsupervised approach for medical concept extraction. In *MedIR Workshop, Sigir*. 1–4.

[126] Graham Spinks and Marie-Francine Moens. 2019. Justifying diagnosis decisions by deep neural networks. *Journal of Biomedical Informatics* 96 (2019), 103248.

[127] J. Springenberg, Alexey Dosovitskiy, Thomas Brox, and M. Riedmiller. 2015. Striving for simplicity: The all convolutional net. In *ICLR (Workshop Track)*.

[128] Li Sun, Weipeng Wang, Jiyun Li, and Jingsheng Lin. 2019. Study on medical image report generation based on improved encoding-decoding method. In *Intelligent Computing Theories and Application*. Springer Intl. Publishing, Cham, 686–696.

[129] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'15)*. 1–9.

[130] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'16)*. 2818–2826.

[131] Jiang Tian, Cong Li, Zhongchao Shi, and Feiyu Xu. 2018. A diagnostic report generator from CT volumes on liver tumor with semi-supervised attention mechanism. In *Medical Image Computing and Computer Assisted Intervention (MICCAI'18)*. Springer Intl. Publishing, Cham, 702–710.

[132] Jiang Tian, Cheng Zhong, Zhongchao Shi, and Feiyu Xu. 2019. Towards automatic diagnosis from multi-modal medical data. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer Intl. Publishing, Cham, 67–74.

[133] Erico Tjoa and Cuntai Guan. 2019. A survey on explainable artificial intelligence (XAI): Towards medical XAI. *arXiv:1907.07374* (2019).

[134] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. 2019. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proc. of the 4th Machine Learning for Healthcare Conference (Proc of Machine Learning Research)*, Vol. 106. PMLR, 359–380.

[135] Eric Topol. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (1st ed.). Basic Books, Inc.

[136] Min-Jen Tsai and Yu-Han Tao. 2019. Machine learning based common radiologist-level pneumonia detection on chest X-rays. In *2019 13th Intl. Conf. on Signal Processing and Communication Systems (ICSPCS'19)*. IEEE, 1–7.

[137] Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proc. of the 14th International Conference on Natural Language Generation*. Association for Computational Linguistics, 140–153. https://aclanthology.org/2021.inlg-1.14.

[138] Emiel van Miltenburg, Wei-Ting Lu, Emiel Krahmer, Albert Gatt, Guanyi Chen, Lin Li, and Kees van Deemter. 2020. Gradations of error severity in automatic image descriptions. In *Proc. of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics, 398–411. https://aclanthology.org/2020.inlg-1.45.

[139] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 5998–6008.

[140] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'15)*. 4566–4575.

[141] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'15)*. 3156–3164.

[142] Jinhua Wang, Xi Yang, Hongmin Cai, Wanchang Tan, Cangzheng Jin, and Li Li. 2016. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Scientific Reports (Nature Publisher Group)* 6 (2016), 27327.

[143] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'17)*. 3462–3471.

[144] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'18)*. 9049–9058.

[145] Xuwen Wang, Yu Zhang, Zhen Guo, and Jiao Li. 2019. A computational framework towards medical image explanation. In *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems*. Springer Intl. Publishing, Cham, 120–131.

[146] Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation* 1, 2 (1989), 270–280.

[147] C. Wu, H. Chang, J. Liu, and J. R. Jang. 2018. Adaptive generation of structured medical report using NER regarding deep learning. In *2018 Conf. on Technologies and Applications of Artificial Intelligence (TAAI'18)*. 10–13.

[148] Luhui Wu, Cheng Wan, Yiquan Wu, and Jiang Liu. 2017. Generative caption for diabetic retinopathy images. In *2017 Intl. Conf. on Security, Pattern Analysis, and Cybernetics (SPAC'17)*. 515–519.

[149] Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, and Shaojie Tang. 2020. A survey on domain knowledge powered deep learning for medical image analysis. *arXiv:2004.12150* (2020).

[150] Xiancheng Xie, Yun Xiong, Philip S. Yu, Kangan Li, Suhua Zhang, and Yangyong Zhu. 2019. Attention-based abnormal-aware fusion network for radiology report generation. In *Database Systems for Advanced Applications*. Springer Intl. Publishing, Cham, 448–452.

[151] Yuxuan Xiong, Bo Du, and Pingkun Yan. 2019. Reinforced transformer for medical image captioning. In *Machine Learning in Medical Imaging*. Springer Intl. Publishing, Cham, 673–680.

[152] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. of the 32nd Intl. Conf. on Intl. Conf. on Machine Learning - Volume 37 (ICML'15)*. JMLR.org, 2048–2057.

[153] Yuan Xue and Xiaolei Huang. 2019. Improved disease classification in chest X-rays with transferred features from report generation. In *Information Processing in Medical Imaging*. Springer Intl. Publishing, Cham, 125–138.

[154] Yuan Xue, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer Antani, George R. Thoma, and Xiaolei Huang. 2018. Multimodal recurrent model with attention for automated radiology report generation. In *Intl. Conf. on Medical Image Computing and Computer-Assisted Intervention*. Springer, 457–466.

[155] C. Yin, B. Qian, J. Wei, X. Li, X. Zhang, Y. Li, and Q. Zheng. 2019. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *2019 IEEE Intl. Conf. on Data Mining (ICDM'19)*. 728–737.

[156] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention (MICCAI'19)*. Springer Intl. Publishing, Cham, 721–729.

[157] Xianhua Zeng, Li Wen, Banggui Liu, and Xiaojun Qi. 2020. Deep learning for ultrasound image caption generation based on object detection. *Neurocomputing* 392 (2020), 132–141.

[158] Xian-Hua Zeng, Bang-Gui Liu, and Meng Zhou. 2018. Understanding and generating ultrasound image description. *Journal of Computer Science and Technology* 33, 5 (2018), 1086–1100.

[159] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV'17)*. 5907–5915.

[160] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proc. of the 28th Intl. Conf. on Neural Information Processing Systems - Volume 1 (NIPS'15)*. MIT Press, Cambridge, MA, 649–657.

[161] Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5108–5120. https://doi.org/10.18653/v1/2020.acl-main.458

[162] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 07 (April 2020), 12910–12917. https://doi.org/10.1609/aaai.v34i07.6989

[163] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proc of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'17)*. 3549–3557.

[164] Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2017. Adversarially regularized autoencoders. *arXiv:cs.LG/1706.04223*.

[165] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'16)*. 2921–2929.

[166] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. 2017. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'17)*. 2027–2036.

[167] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of the IEEE Intl. Conf. on Computer Vision*. 2223–2232.

Received September 2020; revised November 2021; accepted December 2021