MedCycle: Unpaired Medical Report Generation via Cycle-Consistency

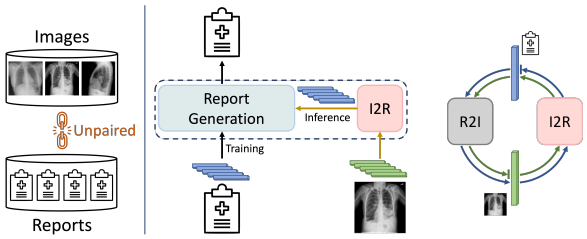
Elad Hirsch **Gefen Dawidowicz** **Ayellet Tal**
Technion – Israel Institute of Technology
{eladhirsch,gefen}@campus.technion.ac.il, ayellet@ee.technion.ac.il

Abstract

Generating medical reports for X-ray images presents a significant challenge, particularly in unpaired scenarios where access to paired image-report data for training is unavailable. Previous works have typically learned a joint embedding space for images and reports, necessitating a specific labeling schema for both. We introduce an innovative approach that eliminates the need for consistent labeling schemas, thereby enhancing data accessibility and enabling the use of incompatible datasets. This approach is based on cycle-consistent mapping functions that transform image embeddings into report embeddings, coupled with report auto-encoding for medical report generation. Our model and objectives consider intricate local details and the overarching semantic context within images and reports. This approach facilitates the learning of effective mapping functions, resulting in the generation of coherent reports. It outperforms state-of-the-art results in unpaired chest X-ray report generation, demonstrating improvements in both language and clinical metrics. Our code is publicly available [1].

1 Introduction

Automating the generation of medical reports has the potential to improve the efficiency of patient information analysis and documentation, leading to better care and cost savings. Consequently, many research efforts have been directed towards this aim (Chen et al., 2021, 2020; Jing et al., 2017; Li et al., 2019; Wang et al., 2022a). These works rely on labeled image-report paired datasets (Demner-Fushman et al., 2016; Johnson et al., 2019), which are relatively small and less accessible in comparison to datasets for natural images (Lin et al., 2014; Sharma et al., 2018; Thomee et al., 2016). Privacy concerns, restricted access to high-quality data, and the complex nature of medical data analysis and labeling, demanding specialized expertise, contribute to this problem and limit the availability of such paired data. Even when paired datasets exist, they may not be fully accessible to the public, leading to partially-available datasets, such as the one presented in (Irvin et al., 2019), which consists solely of images. These limitations in obtaining paired data pose a significant challenge in this domain.



(a) Data (b) Generation via Cycle-Consistent Mapping

Figure 1: **Unpaired medical report generation.** (a) Two unpaired datasets are available: chest X-ray images and chest X-ray reports. (b) Our model learns cycle-consistent mappings between image and report embedding spaces ($I2R$ & $R2I$), facilitated by cross-modality alignment through the use of pseudo-reports, as well as report auto-encoding. Report generation is executed by decoding transformed image representations into reports during inference.

Utilizing unpaired data, i.e., images and reports originating from different sources, may help alleviating some of these limitations. Specifically, it may resolve privacy or regulatory concerns and increase the available amount of training data. Nonetheless, when attempting to learn report generation from images, the absence of image-report pairs introduces a significant challenge. The only previous work that addressed this task involved constructing a knowledge graph of the domain and utilizing a classification module for pathologies (Liu et al., 2021c). This process requires expertise in constructing the knowledge graph and, most importantly, ensuring

[1] https://github.com/eladhi/MedCycle

Findings of the Association for Computational Linguistics: NAACL 2024, pages 1929–1944

that the images and reports are labeled according to a consistent and shared schema. These labels typically pertain to various thoracic pathologies. Our motivation is to eliminate these constraints.

We propose to address the challenge by considering four perspectives, each contributing to a cohesive solution: embedding spaces, mapping between these spaces, initial cross-modality alignment, and report generation (Figure 1). Specifically, we construct two separate embedding spaces—one for the visual modality (images) and the other for the textual modality (reports)—utilizing joint global and local representations. Subsequently, mapping functions learn the transformation from an image representation to a report representation and vice versa. As mapping functions should preserve the semantic meaning of the data, and image-report pairs are unavailable, training a mapping function to transform image representation to its corresponding report representation becomes challenging. To overcome this issue, we train the mapping between the two embedding spaces to preserve cycle consistency. Additionally, to establish initial cross-modality relationships, we introduce a novel concept of pseudo-reports, leveraging available domain information accompanying the image dataset (e.g., pathologies). We encourage the representation of an input image, after mapping to the report space, to closely align with the representation of its corresponding pseudo-report. Lastly, a decoder is exclusively trained with reports, utilizing auto-encoding, to generate medical reports.

Overall, at inference time, given an image, we use both the learned mapping to the report domain and the knowledge learned through auto-encoding to generate a report that suits the image.

Our model adeptly handles a fundamental requirement of report generation—the need for details. Recall that the available data is solely global, indicating the presence of pathologies and encompassing the entire image or report. Alignment at this global level often results in overly generalized representations, which may be suitable for classification but fall short in capturing fine details in individual examples—details crucial for effective report generation. Conversely, local representations—those depicting image patches or individual report words—capture numerous details, providing essential information for the report generation process. However, they lack alignment.

The effectiveness of our method is evident from improvements in both language and clinical metrics. For instance, when compared to previous SoTA methods, our approach demonstrates a 9% enhancement in the BLEU-1 score (language efficacy) and a 3% increase in F1 (clinical efficacy) on the dataset from (Johnson et al., 2019), all while eliminating the need for specific training dataset requirements. Furthermore, we illustrate how the absence of these requirements allows the utilization of other training datasets, resulting in further performance improvements.

Hence, our paper makes the following contributions:

1. We introduce a novel approach to generate medical reports from images in an unpaired setting. This approach is based on learning cycle-consistent mapping functions between domains, establishing cross-modality relations through a novel concept of pseudo-reports, and utilizing an auto-encoding model to generate reports from images.

2. Our method eliminates the need for image and report datasets to be labeled with a consistent schema, thereby increasing data accessibility. This enables the utilization of datasets that were previously incompatible due to differences in pathology labels or languages.

3. Our method outperforms the SoTA results in unpaired chest X-ray report generation.

2 Related Work

Paired medical report generation. Methods that rely on paired data have access to both images and their corresponding reports. These models typically employ an encoder-decoder architecture, where the encoder extracts visual features, commonly using a CNN, and the decoder generates text. Some models utilize a hierarchical decoder comprising topic and word decoders (Jing et al., 2017; Liu et al., 2019; Zhang et al., 2020), while others employ Transformers (Chen et al., 2021, 2020; Huang et al., 2023; Li et al., 2023; Hou et al., 2023). Knowledge graphs (Li et al., 2019; Liu et al., 2021b; Zhang et al., 2020) and memory blocks (Chen et al., 2021, 2020; Wang et al., 2022a,b) are commonly used to learn and encode priori domain information.

Unpaired medical report generation. The only work that addresses the task of unpaired medical report generation is KGAE (Liu et al., 2021c). This

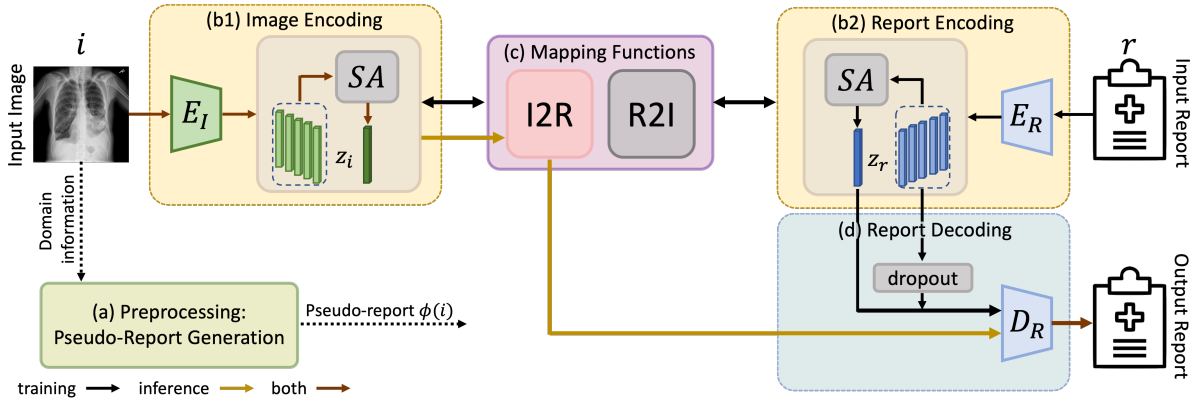


Figure 2: **Method.** For each image i from a dataset $\mathcal{D}_I$, a preprocessing step generates a corresponding pseudo-report denoted as $\phi(i)$ (a), which conveys essential image information in textual form. An image encoder encodes each image i into z_i (b1). Simultaneously, a report encoder encodes reports from a report dataset $\mathcal{D}_R$ (b2), as well as pseudo-reports. These encoded representations comprise both local and aggregated global features, by employing self-attention SA. Two mapping functions are trained to transform image representations into report representations ($I2R$), and vice versa ($R2I$) (c) . Subsequently, a decoder (d) utilizes the encoded reports (excluding pseudo-reports) to output a report, aiming to reconstruct the initial report. For improved generalization, dropout masks a portion of the local representations. During inference, an input image is encoded (b1), followed by mapping to the report space (c). The transformed representation is then decoded to generate a report (d).

work uses a pre-constructed knowledge graph, together with image labels and report labels (regarding several thoractic pathologies), to map images and reports to a shared embedding space. Notably, the image and report datasets must follow the same labeling schema.

Unpaired image captioning. In the domain of natural images, there is an abundance of auxiliary data and pre-trained models available to establish connections between vision and language. Common approaches include object-centric methods that rely on external annotated sources (Hendricks et al., 2016; Venugopalan et al., 2017) and the use of pre-trained models such as object detectors and classifiers (Feng et al., 2019; Gu et al., 2019; Laina et al., 2019; Liu et al., 2021a; Meng et al., 2022). Conversely, in the medical domain, object-based approaches are unsuitable due to the primary focus on diagnosis. Locating abnormalities is challenging due to their size, distribution, relation to other organs, and the limited availability of data.

Unsupervised machine translation. The task of translating text between languages without relying on parallel corpora or human supervision, has also gained attention in recent years. Advancements occurred thanks to initialization schemes and the back-translation approach, which rely on generating pseudo-language pairs (Artetxe et al., 2018; Lample et al., 2018a,b) or extracting them from a real corpus (Wu et al., 2019). Similar approaches have also been employed in unsupervised speech-to-speech translation (Nachmani et al., 2023).

Cycle consistency. The concept of cycle consistency has been explored across various domains, including image-to-image translation (Zhu et al., 2017; Hoffman et al., 2018; Huang et al., 2018) and machine translation (He et al., 2016a), where it is employed to address absence of paired data. Recently, CycleNet (Xu et al., 2024) has also demonstrated the efficacy of this approach in regularizing image manipulation by diffusion models.

3 Method

Our goal is to create a model capable of generating medical reports for X-ray images, using two separate datasets–one for reports and the other for images. The grand challenge lies in the absence of paired image-report data during training; in other words, there is no direct correlation between a report from one dataset and an image from the other. We propose to tackle the problem from four perspectives, each addressing a different aspect of the challenge, which collectively provides a coherent solution. First, we establish cross-modality relationships through a new concept: pseudo-reports. Second, encoding images and reports using similar procedures and relying on joint global and local representations will enable mapping between the embedding spaces. Third, cycle-consistent mapping functions will learn how to transform an image rep-

resentation into a report representation. Finally, by using only reports, a decoder is trained to generate medical reports.

Our method, illustrated in Figure 2, incorporates these four ideas. Its preprocessing procedure generates pseudo-reports. Additionally, the model comprises two encoders: one dedicated to image encoding and the other to report encoding. Furthermore, it integrates two mapping networks: one responsible for converting image representations into report representations ($I2R$), and another for the reverse transformation ($R2I$), thereby implementing the third idea. Finally, the model incorporates a report decoder, encapsulating the fourth idea.

During inference, an input image is encoded by the image encoder. The encoded image representation is then mapped to the report space by a mapping function ($I2R$). Given this transformed representation, the decoder outputs a report that correlates to the input image.

Below, we use the following notations: $\mathcal{D}_I$ is the image datasets, $i \in \mathcal{D}_I$ is an image and z_i is i's representation. The pseudo-report of image i is $\phi(i)$ and its representation is $z_{\phi(i)}$. $\mathcal{D}_R$ is the report datasets, $r \in \mathcal{D}_R$ is a report and z_r is r's representation.

3.1 Preprocessing: Pseudo-report generation

Given an image, the basic challenge lies in establishing a relationship between its representation and a relevant report representation, despite the lack of paired data. We propose a simple approach to solve this problem: generate a *pseudo-report* by employing available domain information. While this report may not be a detailed report, it provides an image-report relationship that shares semantic similarities. We will demonstrate our ability to utilize such pairs for guiding the mapping functions.

The pseudo-reports are pieces of text generated by leveraging domain-specific information available for images. For example, if labels indicating the presence or absence of specific pathologies are accessible, we incorporate this information into the pseudo-reports by describing them with related phrases commonly found within the reports. If reports are available in languages other than English, we rely on an automatic translator, although it may not be optimized for the medical domain, to produce these pseudo-reports. In both cases, inaccuracies in terms of deviations from human-written reports and the level of detail may be introduced. However, these pseudo-reports suffice for our goal of simply maintaining similarity in high-level semantic content, thus highlighting the essence of the data. Additionally, significant errors should not be present, as our pseudo-reports rely on existing data within available datasets.

To understand the differences between a report and a pseudo-report, a report from a study concerning atelectasis and cardiomegaly could be *"Low lung volumes and distended bowel as described on concurrent CT abdomenpelvis. There are patchy opacities suggesting minor dependent bibasilar atelectasis. There is persistent cardiomegaly. There is no pneumothorax or pleural effusion."* Our related pseudo-report could be *"There is cardiomegaly. There is atelectasis. No pleural effusion. No pneumothorax."* Notably, the latter provides partial information and differs in style.

Cross-modality constraint. During model training, these pseudo-reports serve as a cross-modality constraint, to encourage similarity between matching global representations. Given an image and its corresponding pseudo-report, our objective is to ensure that the transformed (i.e. mapped by $I2R$) global representation of an image closely aligns with that of its pseudo-report, and similarly that the transformed (i.e. mapped by $R2I$) global representation of a pseudo-report closely resembles that of the original image. This constraint is implemented by the following loss:

$$
\begin{aligned}
D_{RS} &= \Delta_{cnt}\big(I2R(z_i), z_{\phi(i)}, \{z_{\phi(j)}|_{j\neq i}\}\big) \\
D_{IS} &= \Delta_{cnt}\big(z_i, R2I(z_{\phi(i)}), \{R2I(z_{\phi(j)})|_{j\neq i}\}\big) \\
L_{cm} &= \frac{1}{2|\mathcal{D}_I|} \sum_{i\in\mathcal{D}_I, j\neq i} \Big(D_{RS} + D_{IS}\Big),
\end{aligned}
\tag{1}
$$

where $\Delta_{cnt}(a, b, C)$ measures the dissimilarity between the global representations by employing a contrastive loss, considering a and b a positive pair and C the set of negatives.

3.2 Report & image encoding

Our aim when processing a report or an image is to extract valuable representations that encapsulate the subtleties of the data and can later be decoded into a coherent and informative report. In the training phase, the report representation will serve to reconstruct the input report through auto-encoding. However, during inference, the image representations will be utilized to generate a report.

A key observation in chest X-ray images and reports is that indicators of abnormalities are often

subtle, occupying only a small portion of the image or a few words in the report (Dawidowicz et al., 2023). However, detecting these abnormalities is the core goal in this domain. Hence, we propose to employ two levels of representation: the local level, which focuses on image patches and report words, and the global level, encompassing the entirety of the image and report. While the global representation should capture the essence, such as the presence of pathologies, the local representation delves into a multitude of details that appear in the image or the report.

Typically, encoders extract local features. We propose to generate the global representation by a weighted sum of the local representations, utilizing self-attention (Lin et al., 2017). Hence, the influence of distinct local representations on the global representation varies based on their content and significance. In addition, the connection between global and local representations ensures that any loss imposed on the global representation propagates to the local representations.

3.3 Cycle-consistent mapping

The mapping, *I2R*, is the key component during inference, as it transforms image representations into report representations. Since image-report pair correspondence is unavailable, we propose to train this module using a cycle-consistency constraint. This is done as follows: First, we employ an additional transformation that converts report representations into image representations, $R2I$. Hence, we create a cycle where an image representation is mapped to a report representation by the $I2R$ module and then back by the $R2I$ module. Second, to ensure that $I2R$ outputs valid report representations, we apply these modules in the reversed order as well, starting from a report representation, demanding cycle consistency for that modality too. Finally, to further promote the similarity of distributions of the spaces before and after the transformations, we employ adversarial training. We elaborate hereafter.

Cycle objective. Our approach requires consistency in two cycles: from image to image through report and vice versa. For both cycles, we demand that the reconstructed representation $\hat{z}_i$ ($/\hat{z}_r$) resembles the original representation z_i ($/z_r$). For instance, given an image representation z_i, its reconstruction $\hat{z}_i$ is attained by applying both mapping functions, $I2R$ and $R2I$, sequentially, i.e. $\hat{z}_i = R2I(I2R(z_i))$. Hence, the loss is:

$$\begin{aligned} L_{cyc} = &\frac{1}{|\mathcal{D}_I|} \sum_{i \in \mathcal{D}_I, j \neq i} \Delta_{cnt}(z_i, \hat{z}_i, \{\hat{z}_j\}) + \\ &\frac{1}{|\mathcal{D}_R|} \sum_{r \in \mathcal{D}_R, s \neq r} \Delta_{cnt}(z_r, \hat{z_r}, \{\hat{z_s}\}) \,, \end{aligned} \tag{2}$$

where $\Delta_{cnt}(a, b, C)$ quantifies the dissimilarity between the global representations employing contrastive loss, as in Eq. 1.

Adversarial regularization. To ensure the intended performance of our decoder during inference, its input should resemble the training data. In our case, this implies that the transformed image representations $I2R(z_i)$ should appear as though they were sampled from the report space. Toward this end, we propose to employ adversarial training, which aims to align embedding spaces, making two spaces indistinguishable. For that purpose, we utilize an auxiliary neural network that functions as a discriminator during training (Ganin et al., 2016).

Throughout training, the discriminator's objective is to distinguish between the embedding vectors from the source space (prior to mapping) and the target space (after mapping). Given the representations z_i and z_r, along with their respective mappings $I2R(z_i)$ and $R2I(z_r)$, the discriminator attempts to classify z_i & z_r into one class (a pre-mapping class) and $I2R(z_i)$ & $R2I(z_r)$ into another class (a post-mapping class). The encoders and the mapping modules ($I2R$ and $R2I$) are trained to fool the discriminator, promoting indistinguishable representations. The discriminator produces four predicted probabilities, denoted as $p_{disc}(0|z_r)$ and $p_{disc}(0|z_i)$, representing the probability of the original representation to belong to the pre-mapping space, where $p_{disc}(1|R2I(z_r))$ and $p_{disc}(1|I2R(z_i))$ represent the probability of the transformed representations to belong to the post-mapping space. The discriminator is trained to minimize the following loss function:

$$\begin{aligned} L_{disc}^{(r)} = &-\frac{1}{|\mathcal{D}_R|} \sum_{r \in \mathcal{D}_R} \log\left(p_{disc}(0|z_r)\right) + \\ &-\frac{1}{|\mathcal{D}_I|} \sum_{i \in \mathcal{D}_I} \log\left(p_{disc}(1|I2R(z_i)\right) . \end{aligned} \tag{3}$$

Here, we maximize the likelihood that the discriminator classifies the report representations as belonging to the pre-mapping space (the 0 class) and that the transformed image representation belongs to the post-mapping space (the 1 class). We similarly

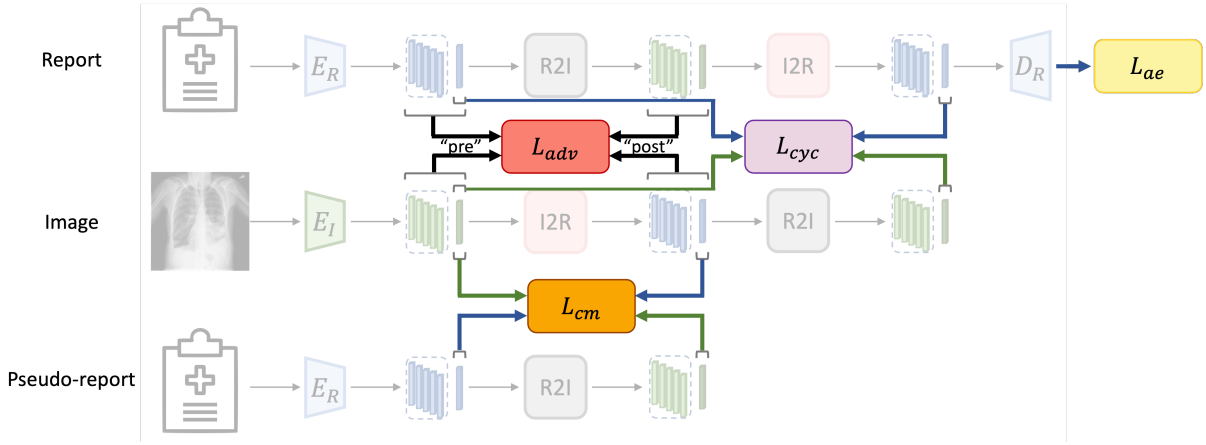


Figure 3: **Training objectives.** Our training involves four distinct objectives, each corresponding to a different loss. The auto-encoding loss (yellow) focuses on accurately reconstructing the input report. The cycle loss (violet) ensures cycle consistency in the $I2R$ and $R2I$ mappings. The adversarial loss (red) ensures that the representations exhibit the same distribution before and after the mapping. Lastly, the cross-modal loss (orange) aims to constrain the mapping by ensuring that pseudo-reports containing information related to an input image and the corresponding image have similar global representations.

compute $L^{(i)}_{disc}$ for discriminating z_i and $R2I(z_i)$. Then, $L_{disc} = L^{(r)}_{disc} + L^{(i)}_{disc}$.

Recall that our model aims to deceive the discriminator. Hence, it has a loss L_{adv} that has the same structure as L_{disc}, but with the labels 0 and 1 swapped. In other words, it aims to make the discriminator classify the representations incorrectly.

3.4 Report decoding

The decoder shall translate latent representations, either of a report or an image, into a coherent report. Due to the lack of image-report pairs, we only train it to generate text from representations of textual reports. The generation process is learned through the auto-encoding of an input report, a task that requires only a dataset of reports.

Nevertheless, the auto-encoding task is prone to overfitting, and such models often learn to copy the input word by word. To increase the generalization capabilities of our decoder, we apply distortions to its input during training by masking out vector representations (input dropout). We apply dropout to the decoder's input for local representations, which express the many details appearing in every report, but not to the global ones which aim to capture high-level semantics.

Auto-encoding objective. The auto-encoding procedure aims to reconstruct input reports. Given a report r, its reconstruction $\hat{r}$ should be as identical as possible to r. For that purpose, we formulate the following loss function:

$$L_{ae} = \frac{1}{|\mathcal{D}_R|} \sum_{r \in \mathcal{D}_R} \Delta_{ce}(r, \hat{r}) \,, \tag{4}$$

where $\Delta_{ce}(\cdot, \cdot)$ measures the dissimilarity between the two reports, calculated as the summation of token-level cross-entropy. In our experiments, we found that a dropout value of probability $p = 0.9$ yielded good results.

Overall loss. The final training objective is the sum of all the previously mentioned objectives (Figure 3). For the report generation model it is:

$$L = \gamma_1 \cdot L_{cm} + \gamma_2 \cdot L_{cyc} + \gamma_3 \cdot L_{adv} + \gamma_4 \cdot L_{ae}. \tag{5}$$

The parameters $\gamma_1, \ldots, \gamma_4$ are hyper-parameter weights. We set them to $\gamma_1 = 3, \gamma_2 = 1, \gamma_3 = 0.25, \gamma_4 = 1.5$. In practice, the loss is computed for a single batch every training iteration.

4 Experimental Results

Datasets. We trained our model using chest X-ray images from the CheXpert dataset (Irvin et al., 2019) or from the PadChest dataset (Bustos et al., 2020), while the training reports were obtained from the MIMIC-CXR dataset (Johnson et al., 2019). For performance evaluation, we utilized test sets from MIMIC-CXR and IU X-ray (Demner-Fushman et al., 2016). Our experimental configuration closely aligns with that of (Liu et al., 2021c), with the distinction that they additionally trained

Method	MIMIC-CXR						IU X-ray					
	B-1	B-2	B-3	B-4	M	R-L	B-1	B-2	B-3	B-4	M	R-L
KGAE CheXpert	0.221	0.144	0.096	**0.062**	0.097	0.208	0.417	0.263	0.181	0.126	0.149	0.318
MedCycle CheXpert	**0.309**	**0.167**	**0.098**	0.061	**0.115**	**0.216**	**0.461**	**0.290**	**0.201**	**0.143**	**0.182**	**0.332**
MedCycle PadChest	0.349	**0.195**	**0.115**	**0.072**	0.128	0.239	**0.479**	**0.291**	0.198	0.140	**0.197**	**0.360**
MedCycle Xlate	**0.352**	0.194	0.114	0.070	**0.132**	**0.241**	0.432	0.277	0.186	0.128	0.188	0.325

Table 1: **Quantitative evaluation, NLG metrics.** Our MedCycle results outperform those of KGAE (Liu et al., 2021c) across all datasets & most of natural language generation metrics: BLEU (B), METEOR (M) & ROUGE-L (R-L), when trained on images from CheXpert dataset. Our results improve when the model is trained on images from PadChest, a dataset that cannot be supported by KGAE. For the IU X-ray dataset, our model was not exposed to any data from the dataset during training, whereas KGAE uses its reports.

on reports from the IU X-ray dataset. We followed the same report preprocessing steps as (Chen et al., 2021; Liu et al., 2021c), which involved filtering out reports lacking a findings section. Importantly, no paired samples were available between the CheXpert (/PadChest) dataset and either the MIMIC-CXR or IU X-ray datasets.

The above datasets are elaborated upon in Appendix C. Generally speaking, each image is CheXpert is assigned with multi-labels for 14 potential diagnosis classes and the corresponding medical reports are not publically available. PadChest is a substantial Spanish dataset of chest X-ray images, associated with medical reports. Each image is labeled according to 174 possible radiographic findings. MIMIC-CXR and IU X-ray are English dataset of chest radiographs, containing images and corresponding reports.

Evaluation metrics. We evaluate our model on two aspects: the quality of the generated language (*NLG*) and its clinical efficacy (*CE*). For NLG evaluation, we employ the BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004) metrics to measure the similarity between the generated reports and the ground-truth. For CE assessment, we utilize the CheXpert (Irvin et al., 2019) model to attribute 14 diagnosis classes related to thoracic diseases and support devices. We then calculate precision, recall & F1 score in comparison to the ground-truth labels.

Quantitative evaluation. Table 1 provides a NLG comparative analysis between our method and KGAE (Liu et al., 2021c), which is the only work addressing the same task. When trained on images from CheXpert, our model outperforms KGAE's across all metrics, except for BLEU-4 on a single dataset where it remains competitive. When trained on images from PadChest, instead of CheXpert, our results improve on the same test datasets (MIMIC-CXR & IU X-ray). This can be explained by the more detailed data on additional pathologies and the availability of Spanish reports. Notably, (Liu et al., 2021c) is unable to utilize PadChest for training, due to its distinct labeling schema compared to the report dataset MIMIC-CXR (174 vs. 14 labels). We explore two approaches for generating pseudo-reports for PadChest: extracting the provided labels (PadChest) or translating the accompanying Spanish reports into English using a general translator (Google Translate; PadChest-Xlate).

Method	Precision	Recall	F1
KGAE CheXpert	0.214	0.158	0.156
MedCycle CheXpert	**0.230**	**0.171**	**0.183**
MedCycle PadChest	**0.237**	0.197	0.183
MedCycle Xlate	0.218	**0.209**	**0.198**

Table 2: **Quantitative evaluation, CE metrics.** Our MedCycle results outperform those of KGAE in terms of clinical efficacy metrics, on MIMIC-CXR, when trained on images from CheXpert or from PadChest.

The improved results demonstrate the potential of leveraging varying datasets, extending beyond those sharing similar labeling schemas. Moreover, our method achieves results for the IU X-ray dataset in a zero-shot manner, implying no exposure to any data from this dataset during training – neither images nor reports. In contrast, (Liu et al., 2021c) utilizes its reports for training.

Our reports not only resemble the ground-truth but also demonstrate higher accuracy and informativeness in extracting clinical information. These findings are depicted in Table 2. As discussed in (Chen et al., 2021), these metrics cannot be employed on IU X-ray dataset, due to its labeling schema, hence Table 2 focuses on MIMIC-

(a) Input image

There are low lung volumes. The lungs are clear. There is no pleural effusion or pneumothorax. The cardiomediastinal silhouette is unremarkable. Left central line terminates in the right atrium. Median sternotomy wires and mediastinal clips are noted. A calcified lymph node is noted in the AP window.

(b) Ground-truth report

Compared to prior exam from. The lungs are clear. There is no pleural effusion or pneumothorax. The cardiomediastinal silhouette is normal. Median sternotomy wires are intact.

(c) Our report

Figure 4: **Qualitative evaluation.** Our model-generated report (c) contains similar information to the ground-truth report (b). It indicates the lung clarity, the cardiomediastinal silhouette's state, the appearance of median sternotomy wires, and rules out pleural effusion & pneumothorax.

CXR. Remarkably, training on PadChest leads to improved results for these clinical metrics as well.

In Appendix B, we compare the unpaired methods to several paired ones. Naturally, SoTA paired methods perform better, but our approach shows promising improvements towards closing the gap.

Qualitative evaluation. The comparison between our generated report and the ground truth is illustrated in Figure 4. Notably, our generated reports contain similar information to what appears in the ground truth, such as the lung clarity, the appearance of sternotomy wires, and absence of pleural effusion or pneumothorax. Additional examples can be found Appendix A. (Making qualitative comparisons to (Liu et al., 2021c) is unfeasible as their code has not been released.)

Pseudo-reports vs. real-reports. To measure the similarity between the pseudo-reports and real-reports, we compared the embeddings of ground-truth reports (from the test set of (Johnson et al., 2019)) against those of pseudo-reports using two metrics recently utilized in (Yu et al., 2023). According to BERTScore (Zhang et al., 2019), the similarity is 0.296, and with CheXbert (Smit et al., 2020), the similarity is 0.407. According to the experiments conducted by (Yu et al., 2023), these scores indicate a good level of similarity.

Implementation details. The encoder E_R is a sequence of an encoding layer (Bengio et al., 2000) and three Transformer encoder layers (Vaswani et al., 2017), while the encoder E_I is a sequence of a ResNet-101 (He et al., 2016b) and three Transformer encoder layers. The decoder D_R is a sequence of three Transformer decoder layers. Both $I2R$ and $R2I$ are implemented as a simple multi-head attention layer with 8 heads. For Δ_{cnt} we set the temperature value to $\tau = 0.1$. We train with a batch size of 128 on a single NVIDIA A100 GPU.

L_{cm}	L_{cyc}	L_{adv}	L_{ae}	B-1	B-4	M	F1
✓	✓	✓	✓	**0.309**	**0.061**	**0.115**	**0.183**
	✓	✓	✓	0.255	0.055	0.103	0.084
✓		✓	✓	0.294	0.060	0.113	0.145
✓	✓		✓	0.286	0.055	0.105	0.151
✓	✓	✓		0.000	0.000	0.003	0.023

Table 3: **Ablation study, losses.** Every loss contributes to the overall improvement in performance across all metrics, including both language and clinical aspects.

Additional details appear in Appendix D.

5 Ablation Study

Losses. Table 3 demonstrates that optimal performance is achieved when combining all our objectives. Applying L_{ae} is crucial for training the generation of reports; otherwise, the decoder fails to learn to produce samples from the report domain. L_{cm} plays a significant role in generating reports closely associated with the input image. The absence of L_{cm} results in relatively poor performance, particularly in the $F1$ metric, suggesting potential inaccuracies in capturing the essential information of the data – the pathologies. Lastly, both L_{cyc} and L_{adv} contribute to further enhancing the results, as they are applied on representations of actual reports and images, rather than the pseudo-reports.

Global & local representations. Table 4 illustrates the impact of employing both global and local representations as inputs for the decoder. Across most metrics, utilizing both representations yields better results. Notably, the F1 score highlights that using only one representation leads to an inadequate expression of essential data elements. Solely relying on the global representation produces favorable results in terms of NLG metrics but at the expense of a significantly poor F1 score. Employing only the local representation enhances the F1 score but

decode w/global	decode w/local	B-1	B-4	M	F1
✓	✓	**0.309**	0.061	**0.115**	**0.183**
✓		0.307	**0.067**	0.111	0.123
	✓	0.288	0.061	0.109	0.152

Table 4: **Ablation study, global & local representations.** Both representations contribute to the decoding process, especially in terms of the F1 metric.

results in a decline in NLG metrics.

6 Conclusions

This paper presents a novel approach to generate X-ray reports in an unpaired manner, eliminating the need for paired images and reports during training. Our method integrates four key components: (1) Learning a mapping function between the image and report spaces through cycle-consistency. (2) Creating representations based on both local and global information that suit the problem and the domain. (3) Learning report auto-encoding. (4) Generating pseudo-reports utilizing domain knowledge associated with the image dataset.

We show the effectiveness of our method on two different datasets, surpassing the performance of existing unpaired techniques for generating chest X-ray reports. For instance, when trained on the same image dataset as previous methods, our approach improves the BLEU-1 score (language metric) by 4%-9%, depending on the dataset and the F1 score (clinical metric) by 3%. When trained on different datasets, which could not be utilized by other unpaired methods due to distinct labeling schemas, the results are further improved.

While our method exhibits generality, future availability of sufficiently comprehensive datasets in other medical domains may broaden the scope of this work to encompass other types of medical data.

Limitations. Our model requires the image dataset to be accompanied by relevant domain-specific information, such as pathologies or reports in some language. There are datasets that lack such information and, as a result, cannot be utilized. Furthermore, our model generates a report based on a single input image. However, medical examinations often reference previous findings or compare changes in severity over time, information that might be available from another image or a summary report. Our model is unable to utilize this contextual information despite its significance.

Ethical considerations. In the medical domain data privacy is a core concern. The datasets we employed (Demner-Fushman et al., 2016; Johnson et al., 2019; Bustos et al., 2020) were de-identified and anonymized to ensure privacy protection, in compliance with the Health Insurance Portability and Accountability Act (HIPAA). However, deploying such models on private datasets (e.g., hospital archives) without robust privacy measures may risk the exposure of personal and sensitive information through generated reports. Notably, compared to paired approaches, our unpaired methodology enhances privacy by not relying on paired patient data, mitigating potential privacy breaches. We hope that when unpaired models, such as ours, demonstrate their performance potential, the opportunities that arise with fewer restrictions would encourage data owners (e.g., hospitals) to release such unpaired data, which involves fewer ethical concerns.

In terms of application, automatic report generation aims to enhance patient care and alleviate the burden on healthcare providers. Nevertheless, the automated system remains susceptible to errors, which could result in inaccurate diagnoses. Considering the profound consequences of erroneous diagnoses, we advocate that such automated systems should complement radiologists rather than replace them in real-world applications.

Acknowledgements. We gratefully acknowledge the support of the Israel Science Foundation 2329/22.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

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

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Gefen Dawidowicz, Elad Hirsch, and Ayellet Tal. 2023. Limitr: Leveraging local information for medical image-text representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21165–21173.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. 2019. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10323–10332.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016a. Dual learning for machine translation. *Advances in neural information processing systems*, 29.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023. ORGAN: Observation-guided radiology report generation via tree reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8108–8122, Toronto, Canada. Association for Computational Linguistics.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

Alistair EW 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):317.

Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7414–7424.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049. Association for Computational Linguistics.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6666–6673.

Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2863–2874.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *5th International Conference on Learning Representations, ICLR*.

Fenglin Liu, Meng Gao, Tianhao Zhang, and Yuexian Zou. 2021a. Exploring semantic relationships for unpaired image captioning. *arXiv preprint arXiv:2106.10658*.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762.

Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. 2021c. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, 34:16266–16279.

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*, pages 249–269. PMLR.

Zihang Meng, David Yang, Xuefei Cao, Ashish Shah, and Ser-Nam Lim. 2022. Object-centric unsupervised image captioning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 219–235. Springer.

Eliya Nachmani, Alon Levkovitch, Yifan Ding, Chulayutsh Asawaroengchai, Heiga Zen, and Michelle Tadmor Ramanovich. 2023. Translatotron 3: Speech to speech translation with monolingual data. *arXiv preprint arXiv:2305.17547*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning images with diverse objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5753–5761.

Jun Wang, Abhir Bhalerao, and Yulan He. 2022a. Cross-modal prototype driven network for radiology report generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 563–579. Springer.

Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. 2022b. A medical semantic-assisted transformer for radiographic report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022*, pages 655–664. Springer.

Jiawei Wu, Xin Wang, and William Yang Wang. 2019. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

1 (Long and Short Papers), pages 1173–1183. Association for Computational Linguistics.

Sihan Xu, Ziqiao Ma, Yidong Huang, Honglak Lee, and Joyce Chai. 2024. Cyclenet: Rethinking cycle consistency in text-guided diffusion for image manipulation. *Advances in Neural Information Processing Systems*, 36.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12910–12917.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

A Additional Examples

Matching phrases in each ground-truth and generated report pair are marked with the same color.

(a) Input image

As compared to the previous radiograph there is no relevant change. No evidence of pneumonia. Borderline size of the cardiac silhouette without pulmonary edema. No pleural effusions. No inhomogeneous bone structure. Mild tortuosity of the thoracic aorta.

(b) Ground-truth report

As compared to the previous radiograph no relevant change is seen. Borderline size of the cardiac silhouette with tortuosity of the descending aorta. No pulmonary edema.

(c) Our report

(a) Input image

Frontal and lateral views of the chest were obtained. The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. Cardiac and mediastinal silhouettes are unremarkable.

(b) Ground-truth report

PA and lateral views of the chest. The lungs are clear without focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal.

(c) Our report

(a) Input image

Single frontal view of the chest was obtained. A left pleural effusion with overlying atelectasis remains present. Left base retrocardiac opacity likely represents combination of atelectasis and effusion although underlying consolidation is difficult to exclude. Patient is status post median sternotomy and CABG. No definite focal consolidation is seen in the right lung. The patient is status post median sternotomy and cardiac valve replacement. Cardiac and mediastinal silhouettes are stable.

(b) Ground-truth report

The patient is status post median sternotomy and CABG. Left chest tube has been removed. There is a small left pleural effusion with adjacent atelectasis. The heart size is normal. The mediastinal and hilar contours are unremarkable. Left basilar atelectasis is noted. No pneumothorax.

(c) Our report

(a) Input image

As compared to the previous radiograph the right picc line is in unchanged position. Unchanged evidence of mild fluid overload and retrocardiac atelectasis. No overt pulmonary edema. No pneumonia. Moderate cardiomegaly.

(b) Ground-truth report

As compared to the previous radiograph there is no relevant change. Moderate cardiomegaly is stable. There is a small left pleural effusion with compressive atelectasis at the left base. No evidence of pneumothorax or pulmonary edema.

(c) Our report

(a) Input image

PA and lateral views of the chest provided. There is no focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen.

(b) Ground-truth report

PA and lateral views of the chest were obtained. The lungs are clear without focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is within normal limits. No acute osseous abnormality is seen.

(c) Our report

(a) Input image

In comparison with the study of there is continued enlargement of the cardiac silhouette with congestive failure. Poor definition of the hemidiaphragms is consistent with bilateral pleural effusion and compressive atelectasis. There is an area of more coalescent opacification in the right upper zone that is asymmetric with the opposite side. In the appropriate clinical setting this could well represent a developing focus of pneumonia.

(b) Ground-truth report

In comparison with the study of there has been interval improvement in the right pleural effusion with compressive atelectasis. Bibasilar opacities have improved. The heart size remains mildly enlarged. Mediastinal contours are stable. There is small bilateral pleural effusion. No pneumothorax.

(c) Our report

(a) Input image

Patient is status post median sternotomy. Left-sided pacer device is grossly stable in position. There is a moderate left pleural effusion with overlying atelectasis left base consolidation is not excluded. Similar pulmonary edema persists possibly asymmetric on the left. No right pleural effusion is seen. There is no pneumothorax. Cardiac and mediastinal silhouettes are stable.

(b) Ground-truth report

The patient is status post median sternotomy and CABG. There is a small left pleural effusion with adjacent atelectasis in the left lower lobe. The lungs are clear. The heart size is normal. No pneumothorax. A left picc line ends in the mid SVC. No acute osseous abnormalities.

(c) Our report

(a) Input image

The lung fields are clear without focal consolidation pleural effusion or pneumothorax. Heart and mediastinal contours are within normal limits. Sternal wires and mitral valve replacement hardware are again seen.

(b) Ground-truth report

The heart is normal in size. The mediastinal and hilar contours are unremarkable. There is no pneumothorax or pleural effusion. The lungs are clear.

(c) Our report

B Comparison to Paired Methods

Method	Unpaired	B-1	B-4	M	R-L	P	R	F1
R2Gen-CMN (Chen et al., 2021)	✗	0.353	0.106	0.142	0.278	0.334	0.275	0.278
KGAE-Sup (Liu et al., 2021c)	✗	0.369	0.118	0.153	**0.295**	**0.389**	**0.362**	**0.355**
MSAT (Wang et al., 2022b)	✗	0.373	**0.120**	0.143	0.282	-	-	-
KiUT (Huang et al., 2023)	✗	**0.393**	0.113	**0.160**	0.285	0.371	0.318	0.321
UAR (Li et al., 2023)	✗	0.363	0.107	0.157	0.289	-	-	-
KGAE CheXpert (Liu et al., 2021c)	✓	0.221	0.062	0.097	0.208	0.214	0.158	0.156
MedCycle CheXpert (Ours)	✓	0.309	0.061	0.115	0.216	0.230	0.171	0.183
MedCycle PadChest (Ours)	✓	0.349	0.072	0.128	0.239	0.237	0.197	0.183
MedCycle PadChest-Xlate (Ours)	✓	0.352	0.070	0.132	0.241	0.218	0.209	0.198

Table 5: **Comparison to paired methods.** The difference in performance between several paired and unpaired methods is reduced with our MedCycle, compared to KGAE, while also removing additional requirements regarding the training datasets. As expected, SoTA paired methods perform better (on the MIMIC-CXR dataset). All the paired methods were trained and evaluated using images from the MIMIC-CXR datasets, alongside their corresponding reports. Our method and KGAE were not exposed to images from this dataset, but rather images from CheXpert or PadChest and reports from MIMIC-CXR. Solely the evaluation was performed using images from the MIMIC-CXR dataset.

C Datasets

CheXpert (Irvin et al., 2019) is a dataset of chest X-ray images, containing $224,316$ radiographs from $65,240$ patients, collected at Stanford Hospital. Each image is assigned with multi-labels for 14 potential diagnosis classes. The corresponding medical reports are not publically available.

PadChest (Bustos et al., 2020) is a substantial Spanish dataset of chest X-ray images, comprising $160,868$ radiographs from $69,882$ patients, along with their associated medical reports. The data was collected at San Juan Hospital. Each image is labeled according to 174 possible radiographic findings, 19 diagnoses, and 104 anatomic locations. Each image is associated with a report in Spanish.

MIMIC-CXR (Johnson et al., 2019) is a large English dataset of chest radiographs, containing $377,110$ images, corresponding to $227,835$ reports performed at the Beth Israel Deaconess Medical Center. The dataset is split into $368,960$ images ($222,758$ reports) for training, $2,991$ images ($1,808$ reports) for validation, and $5,159$ images ($3,269$ reports) for testing.

IU X-ray (Demner-Fushman et al., 2016) comprises $7,470$ chest X-ray images, each associated with one of $3,955$ reports. We employ the same train-validation-test split of 70%-10%-20% as defined by (Li et al., 2018).

Access to the datasets is granted directly by the dataset owners upon registration and approval, owing to their sensitivity.

D Implementation Details

1. Encoding:
   (a) Visual feature extraction: we use ResNet-101 pretrained on ImageNet, which yields a 7×7 grid of patch features. Images are first resized, such that the smaller edge of the image is of size 256 pixels. Then, the image is cropped to 224×224 pixels.
   (b) After visual feature extraction, we apply a Transformer encoder module, comprises 3 encoding layers with multi-head attention mechanism of 8 heads.
   (c) Textual feature extraction: we use an embedding layer (LUT).
   (d) After textual feature extraction, we apply a Transformer encoder module, comprises 3 encoding layers with multi-head attention mechanism of 8 heads.
   (e) embedding dimension is set to $d = 512$.

2. Decoding:
   (a) D_R is a Transformer decoder, comprises 3 decoding layers with multi-head attention mechanism of 8 heads.
   (b) For augmenting the input of D_R we use a dropout of 0.9.
3. Mapping:
   (a) Both $I2R$ and $R2I$ are implemented as a simple multi-head attention layer with 8 heads.
4. Losses:
   (a) For the contrastive term Δ_{cnt}, we set the temperature $\tau = 0.1$.
5. Optimization parameters:
   (a) We set the loss weights to $\gamma_1 = 3, \gamma_2 = 1, \gamma_3 = 0.25, \gamma_4 = 1.5$. These parameters were empirically derived through experimentation with several combinations.
   (b) We use Adam optimizer, with betas of $(0.9, 0.98)$ and weight decay of $5 \cdot 10^{-5}$.
   (c) The learning rate is set at 10^{-4} for most components, except for the discriminator, which is set at $2 \cdot 10^{-4}$.
   (d) The batch size is set at 128.
6. Inference parameters:
   (a) The reports are samples using beam search, with a beam of size 3.
7. Resources:
   (a) We train the model on a single Nvidia A100 GPU for 6 hours.