Supplementary Materials

Anonymous Authors

1 WHERE DOES THE FORENSIC EXPERT'S EYE FOCUS IN THE CXR AND VRT IMAGES?

CXRs contain more anatomical structure overlaps while VRT images only contain skeletons. Manual methods comparing these imaging materials from antemortem and postmortem [2,13,18,21,42] are usually focused on the overall skeletal morphology and boundaries, as labeled in Figure 1.

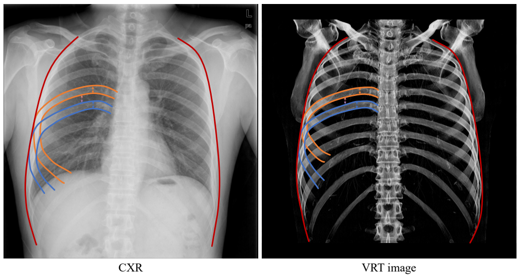

Figure 1: When we compare the CXR with the VRT image from the same individual, we focus on the overall skeletal morphology (red), the boundaries(orange and blue), the width of ribs (orange and blue bidirectional arrows), and the inter-rib space (orange to blue gradient bidirectional arrows).

For the same person, some skeletal details are strikingly similar, the boundaries and curvatures of each rib in the VRT image are almost identical to those in the CXR. However, the whole structure is slightly different. This is because the VRT image and the CXR not only belong to different domains but are also taken in different postures and at different times. When patients take the CXR, they stand up and hold their arms flat with normal breathing. During the VRT image, they lie flat on the plate and raise their arms upward with deep breathing. The different states lead to the deformation of the whole structure of the thoracic skeletons, as shown in Figure 2, which makes the thoracic skeletons in CXRs and VRT images not strictly pixel-level mapping but potentially non-linear stochastic mapping.

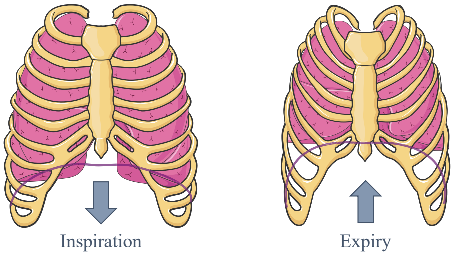

Figure 2: The different respiratory states lead to the deformation of the thoracic skeletons.

2 IMPLEMENTATION DETAILS

Preliminary Since this is a new task, to avoid the interference of automated object detection tools on the region of interest (ROI) of the data, the ROI containing thoracic skeletons was manually cropped from CXRs and VRT images for localization.

Cropping criteria: The area surrounded by the horizontal tangent line of the upper edge of the first thoracic vertebra, the horizontal tangent line of the lower edge of the 12th thoracic vertebra, and the horizontal tangent line of the most lateral edge of the left and right ribs.

Network Architecture For the translation, we use 9 residual blocks for 256×256 images. For the latent fusion, we utilize U-net as the VRT and CXR encoder-decoder modules. The scale of latent features is $1024 \times 16 \times 16$. The latent fusion module is a fully connected layer with $2 \times 1024 \times 16 \times 16$ input nodes and 256 output nodes.

Training Details For the cross-modality translation, we utilize the Adam optimizer with an initial learning rate of 0.0002 and a momentum term of 0.5. The batch size is set to 1. We keep the same learning rate for the first 400 epochs and linearly decay it to zero over the next 100 epochs. For the latent fusion, we employ an SGD optimizer with a learning rate of 0.001 and a momentum of 0.9 and set the batch size to 2. The total epoch number is 300 and the learning rate decays to a tenth of its original value after every 50 epochs.

Data Augmentation To enhance the generalization and robustness of the model, we preprocess all of the input images for data augmentation. All images are resized to 256×256 with the bilinear interpolation. Color jittering involves changing the contrast of the image with a parameter set to 1.8. The probability of color inversion was 0.2. The image is flipped horizontally with a random probability of 0.5. To simulate the different rotation postures of the chest, it is randomly rotated (-20, +20) degrees with a random probability of 0.3. In addition, VRT and translated VRT images are normalized with a mean of 0.3817 and a standard deviation of 0.3180; real and translated CXRs are normalized with a mean of 0.6425 and a standard deviation of 0.1613.

Alternatives Since this is the first comprehensive work in this new task, there is no directly comparable thoracic skeleton identification approach. We employ a cross-modal aligned-based person re-identification method, LbA [38], and several classical baselines for comparison.

We construct three contrastive person identification models based on the Triplet architecture with ResNet-18 as their backbone, as shown in Figure 3(a)-3(c). In Figure 3(a), we only use real samples and extract features for fine-grained skeletal representations based on contrastive learning loss (e.g. InfoNCE [22]). We use the VRT image as the anchor, the CXR from the same person as the positive sample, and the other's CXR as the negative sample. In Figure 3(b)-3(c), we introduce the results of the proposed cross-modality translation and extract features from each image for identifiable representations. We use the real CXR (VRT) image

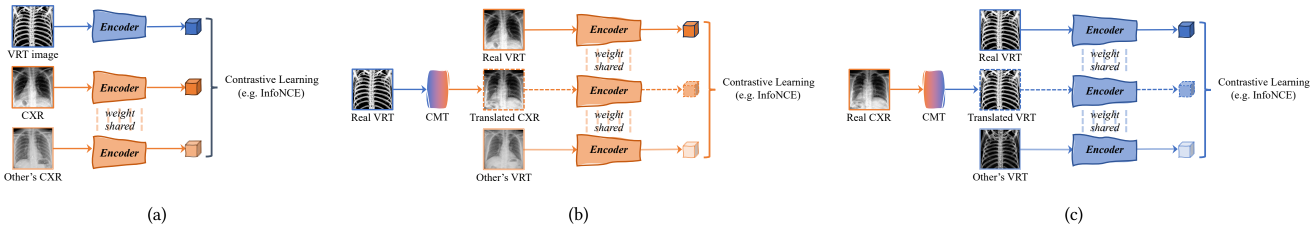

Figure 3: (a) Directly leveraging VRT images and CXRs for identification. (b) Solely identification in the CXR modality based on the translated CXR and the real ones. (c) Solely identification in the VRT modality leveraging the translated VRT images from the real CXRs.

Methods	Rank-k Rate (%) ↑															
	|NBP-Bank|=263 (+0)				+1k distractors				+5k distractors				+10k distractors			
	k = 1	k = 10	k = 20	k = 50	k = 1	k = 10	k = 20	k = 50	k = 1	k = 10	k = 20	k = 50	k = 1	k = 10	k = 20	k = 50
Triplet [24]	0.38	4.94	12.55	27.76	0.38	4.18	7.98	20.53	0.38	3.80	7.60	17.87	0.38	3.80	7.22	17.11
MobileFaceNet [9]	3.42	19.39	30.80	44.49	2.28	10.65	19.39	30.04	2.28	8.75	11.03	18.63	2.28	7.22	8.75	14.07
IResNet-18 [15]	3.04	8.75	14.83	36.50	1.14	2.66	5.70	9.13	1.14	2.28	2.66	4.94	0.76	1.90	2.28	4.18
IResNet-50 [15]	0.76	8.75	12.93	28.90	0.38	3.42	7.98	12.17	0.00	1.52	2.66	5.70	0.00	1.14	1.90	3.80
LbA [38]	0.76	4.94	8.75	24.33	0.76	4.18	8.75	23.57	0.76	4.56	8.75	25.10	0.38	4.18	8.37	24.33
Ours (CMF. + CMT.)	21.29	64.64	79.09	90.49	11.03	36.50	49.05	63.88	8.37	26.62	33.84	46.39	6.84	21.67	27.76	39.92
Triplet [24] + CMT. (VRT)	5.70	27.76	39.54	62.36	2.28	8.37	13.69	27.00	1.14	4.56	6.08	9.51	1.14	3.42	4.94	6.46
Triplet [24] + CMT. (CXR)	2.66	16.73	28.90	52.85	0.38	3.04	5.32	11.79	0.00	0.76	1.14	3.80	0.00	0.38	1.14	1.52
MobileFaceNet [9] + CMT.	11.79	38.78	52.85	68.06	8.37	23.19	29.66	40.68	4.94	14.07	18.63	26.24	3.42	12.17	15.21	20.91
IResNet-18 [15] + CMT.	3.80	14.45	20.15	35.74	1.14	8.37	12.17	15.59	0.38	3.42	5.32	10.27	0.38	2.28	3.42	5.32
IResNet-50 [15] + CMT.	2.66	10.27	13.31	27.38	1.52	7.22	9.89	15.21	0.38	3.42	5.70	7.60	0.38	1.52	3.80	6.08

Table 1: Experiments on gradually introducing 0-10k distractors. Top: Comparison of various approaches. Bottom: Comparison of various approaches with CMT. results introduced, where CMT. represents the cross-modality translation step and CMF. represents the cross-modality fusion step.

Methods		Rank-k Rate (%) ↑															
		|NBP-Bank|=263 (+0)				+1k distractors				+5k distractors				+10k distractors			
λ_{re}	λ_{cl}	k = 1	k = 10	k = 20	k = 50	k = 1	k = 10	k = 20	k = 50	k = 1	k = 10	k = 20	k = 50	k = 1	k = 10	k = 20	k = 50
1	0	1.52	9.51	13.69	33.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	1	22.43	64.26	76.43	90.11	11.03	38.78	47.91	63.50	7.60	26.24	33.08	46.01	6.46	20.53	27.76	37.26
1	1	19.77	64.26	77.95	89.73	9.89	35.36	44.49	60.84	6.08	25.10	31.18	39.92	4.94	17.87	25.86	34.60
1	3	21.29	64.64	79.09	90.49	11.03	36.50	49.05	63.88	8.37	26.62	33.84	46.39	6.84	21.67	27.76	39.92
1	5	16.73	60.08	74.90	90.11	8.75	35.36	44.49	61.22	6.46	23.95	31.18	43.35	6.46	20.15	25.48	36.12

Table 2: Ablation study on the weights for the reconstruction loss and the contrastive learning loss. λ_{re} denotes the weight for the reconstruction loss and λ_{cl} denotes the weight for the contrastive learning loss.

as the anchor, the CXR (VRT) image translated from the real VRT (CXR) of the same person as the positive sample, and the other's real CXR (VRT) image as the negative sample.

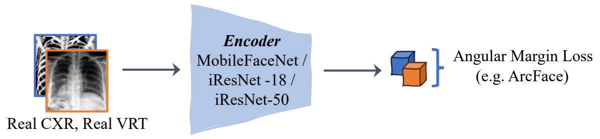

Figure 4: Leveraging the angular margin loss from a classification point of view.

Since the face identification task also aims to learn an identity representation, similar to ours, we also compare our approach with some classical face identification frameworks. As shown in Figure 4, we consider this task as a classification task, just like face classification. In this manner, we employ various models [9, 15] as the feature extractor and utilize a classical face classification loss, ArcFace [14], for comparison.

3 LARGE-SCALE EXPERIMENTS

In this section, we compare our methods with others based on a larger CXR dataset, ChestX-ray8 [1], in Table 1. Specifically, we gradually introduced 0-10k CXRs from ChestX-ray8 as identifying distractors. We set the batch size to 6 with a total of 150 epochs

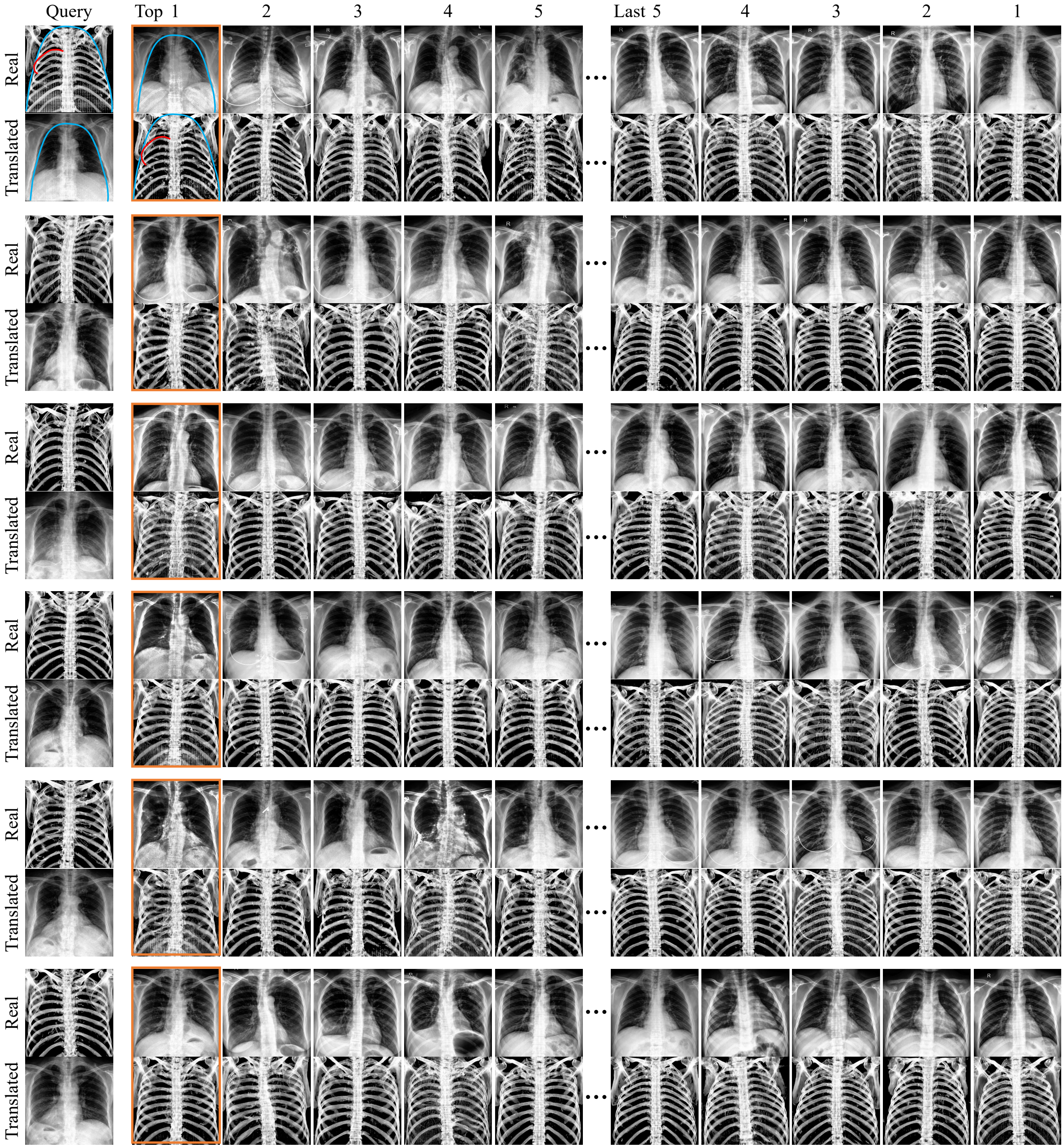

Figure 5: Complementary visualization of large-scale identification. The query pair consists of the real VRT image and its translated CXR. The candidate pairs consist of real CXRs with their translated VRT images. The orange rectangle represents the ground truth. The red and blue lines describe the rib boundary and the overall skeletal morphology respectively. The top-5 identification results are similar to the query one while the last are not.

in the CMF training phase in experiments in this section. Besides, we explore the ablation study on the reconstruction loss and the contrastive learning loss, as shown in Table 2. It shows that both the reconstruction loss and the contrastive loss are necessary, but the weights are required to be properly considered. Directly introducing the reconstruction loss with the same/too small weight as the contrastive loss may lead to a decrease in accuracy.

4 COMPLEMENTARY VISUALIZATION

We show more complementary visualization results in Figure 5. It clearly shows that the top five identification results are highly visually similar to the query while the last five results are significantly different.

REFERENCES

[1] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2097–2106.