Invertible Residual Rescaling Models reviewers' comments：

- Paper Summary:

This paper applies Invertible Rescaling Networks (IRNs) to address the problem of Image Rescaling. They establish long skip connections within the proposed Residual Downscaling Module and short connections within the Invertible Residual Blocks inside RDM. Based on the Invertible Rescaling Networks (IRN) [33], long skip/short connections allow rich low-frequency information to be bypassed by skip connections and force models to focus on extracting high-frequency information from the image. Additionally, IRRM achieves better results than other state-of-the-art methods with much fewer parameters and complexity.

- Paper Strengths:

1. IRRM achieves superior performance compared to other state-of-the-art methods while maintaining much lower complexity and fewer parameters.
2. Long skip connections allow the model to bypass rich low-frequency information, while short skip connections reduce model degradation.
3. The paper provides quantitative comparisons for IRRM on different model sizes and analyzes the impact of residual connections in RDM on training. It explains how residual connections solve the problems of gradient vanishing and explosion.
4. Diffusion index and local attribution maps are analyzed to provide a more intuitive understanding of the effectiveness of IRRM.

- Paper Weaknesses:

1. The technical contributions in this paper are highly similar to those in the IRN paper, only with the addition of long skip connections and short connections and the use of deterministic residual blocks(EB) instead of arbitrary transformation functions.
2. In the "Related Works" section, Line #179-182, there is confusion regarding the phrases "learned separately" in IRN and "removed high-frequency component" in HCFlow. These expressions are not present in the original text.
3. In Figure 3, it appears that the loss of IRRM_Res_PCB decreases more smoothly than that of IRRM_Res_RB, and Table 4 shows that the performance of PCB is comparable(close) to that of RB. This may suggest that the RB settings do not have a significant impact compared to PCB.
4. Since the model settings are similar to those of IRN, it would be useful to compare the comprehensive experimental results of the two models to highlight the technical novelty of IRRM, such as why the use of RB instead of arbitrary transformation functions.
5. There is an error in equation (2) in Section 2.3 concerning the backward process function. It should have a minus sign instead of a plus sign, as compared to the IRN paper.
6. There is no explanation provided for D and A in function (5), and the first introduction of -S, -M, and -L in Table 1 lacks detailed captions or explanations.
7. There is a question about the possibility of long skip connections spanning RDMs for more effective extraction of high-frequency information.

- Overall Recommendation : weak reject
- Confidence Level : 4 The reviewer is confident but not absolutely certain that the evaluation is correct.

For CLIP-ReID：

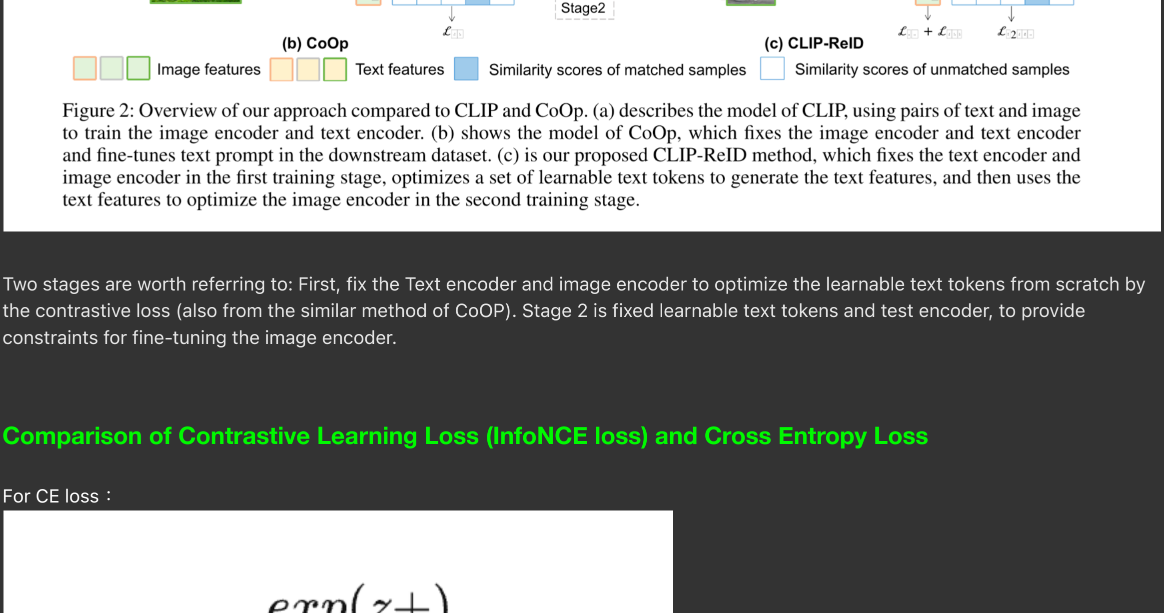


Figure 2: Overview of our approach compared to CLIP and CoOp. (a) describes the model of CLIP, using pairs of text and image to train the image encoder and text encoder. (b) shows the model of CoOp, which fixes the image encoder and text encoder and fine-tunes text prompt in the downstream dataset. (c) is our proposed CLIP-ReID method, which fixes the text encoder and image encoder in the first training stage, optimizes a set of learnable text tokens to generate the text features, and then uses the text features to optimize the image encoder in the second training stage.

Two stages are worth referring to: First, fix the Text encoder and image encoder to optimize the learnable text tokens from scratch by the contrastive loss (also from the similar method of CoOP). Stage 2 is fixed learnable text tokens and test encoder, to provide constraints for fine-tuning the image encoder.

Comparison of Contrastive Learning Loss (InfoNCE loss) and Cross Entropy Loss

For CE loss：

$$-log\frac{exp(z+)}{\sum_{i=0}^{k} exp(z_i)}$$

Since each picture is a separate category, the softmax operation is very time-consuming to calculate on so many categories, coupled with the exponential operation, when the dimension of the vector is several million, the computational complexity is quite high. Therefore, it is not feasible to use CE to calculate loss in contrastive learning.

And for infoNCE loss:

$$L_q = -log\frac{exp(q \cdot k_+/\tau)}{\sum_{i=0}^{k} exp(q \cdot k_i/\tau))}$$

Info NCE loss is a simple variant of NCE. It thinks that if the problem is only regarded as a binary classification, with only data samples and noise samples, it may not be friendly to model learning, because many noise samples may not be a class at all. Therefore, it is more reasonable to regard it as a multi-classification problem. InfoNCE loss is actually a cross entropy loss, which is a classification task of class k+1. The purpose is to classify the picture p into the class k+1.

The smaller the temperature coefficient, the more the model focuses on separating those negative samples that are most similar to this sample

Some thoughts on listening to ChatGPT lectures:

Purpose: To understand human language, to speak human language
ChatGPT=GPT- Generative + Pre-trained + Transformer
Unsupervised training with large corpus of unlabeled languages. Predict the form of the next word for self-learning. -- Since ChatGPT can plan rewards by itself, can it be combined with the current RL model to solve some CV problems?
Language communication can reveal the intelligence of the model - Turing test
The generation is very conservative and safe, what should I do, it takes a long time - Static function vocabulary + dynamically generated content words
If ChatGPT can annotate pictures, all of them can be multimodal unsupervised training — generate reliable pseudo-labels.
Four elements of ChatGPT's success: large model GPT3.5 + reinforcement learning for human feedback + large-scale high-quality prompt data + user feedback to form a flywheel
ChatGPT defects: 1. Factual errors 2. Logical errors 3. Inappropriate remarks 4. Security

Problems encountered when writing related works

- Existing unsupervised methods: How to say the disadvantages of Homogeneous-to-Heterogeneous.

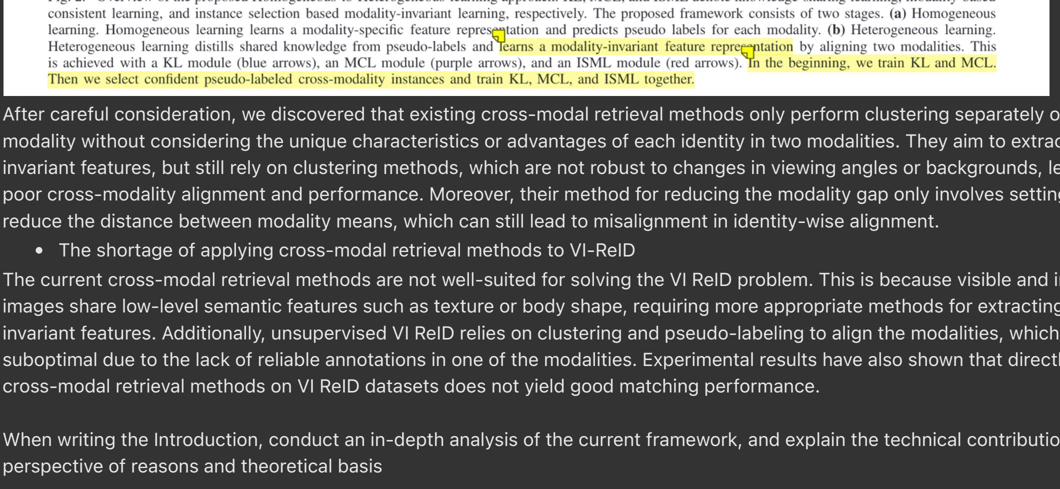


Fig. 2. Overview of the proposed Homogeneous-to-Heterogeneous learning approach. KL, MCL, and ISML denote knowledge-sharing learning, modality-based consistent learning, and instance selection based modality-invariant learning, respectively. The proposed framework consists of two stages. **(a)** Homogeneous learning. Homogeneous learning learns a modality-specific feature representation and predicts pseudo labels for each modality. **(b)** Heterogeneous learning. Heterogeneous learning distills shared knowledge from pseudo-labels and learns a modality-invariant feature representation by aligning two modalities. This is achieved with a KL module (blue arrows), an MCL module (purple arrows), and an ISML module (red arrows). In the beginning, we train KL and MCL. Then we select confident pseudo-labeled cross-modality instances and train KL, MCL, and ISML together.

After careful consideration, we discovered that existing cross-modal retrieval methods only perform clustering separately on each modality without considering the unique characteristics or advantages of each identity in two modalities. They aim to extract view-invariant features, but still rely on clustering methods, which are not robust to changes in viewing angles or backgrounds, leading to poor cross-modality alignment and performance. Moreover, their method for reducing the modality gap only involves setting loss to reduce the distance between modality means, which can still lead to misalignment in identity-wise alignment.

- The shortage of applying cross-modal retrieval methods to VI-ReID

The current cross-modal retrieval methods are not well-suited for solving the VI ReID problem. This is because visible and infrared images share low-level semantic features such as texture or body shape, requiring more appropriate methods for extracting modality-invariant features. Additionally, unsupervised VI ReID relies on clustering and pseudo-labeling to align the modalities, which can be suboptimal due to the lack of reliable annotations in one of the modalities. Experimental results have also shown that directly using cross-modal retrieval methods on VI ReID datasets does not yield good matching performance.

When writing the Introduction, conduct an in-depth analysis of the current framework, and explain the technical contribution from the perspective of reasons and theoretical basis

Problems encountered:

* There is a problem with the PBC module: *the current code does not reflect the function of PBC, it just divides the Identity into n segments, and directly concatenates it, so it has no effect if it is not cut into n segments.*
Initiated thinking: Give a prior: the identity of the same view is divided into 3 pieces, and the size and position of each piece are the same.

So I can cluster by comparing the **ratio of the size** of the three sequential blocks, and the ratio of the **total value of pixels contained in each block.**

But after thinking, this is a very naive solution. For the same identity under different viewing angles, there will be different ratios of the total value of pixels. So this method cannot solve the impact caused by the perspective change (how to consider the view-invariant feature on the basis of using the PBC module when intra-modality clustering?).

Therefore, a further solution is proposed. For the three cut parts, can we find a few points in the distribution center (or a few points around the central area), and form a **topological map** of these points in each part. The role of identity-discriminative feature is played by topology map matching. But this will also increase the complexity of the model, and the point selection problem of the topology map, and how to select points to solve the problem of view shift, so abandon this solution.

finally! Come up with a feasible method, as shown in the following figure:

For learning ID-Discriminative feature.
PBC的作用：根据不同ID(不同view)对foreground进行切分(assume 3块)
n×h×w×c ⟶ PBC module ⟶ 头 上身 下身
对于一个batch，每个sample ID可知，假设为 $S_1, S_2, \cdots, S_n$，则通过PBC，得到 $S_1^{头}, S_1^{上}, S_1^{下}$ 类似的label. 对一个batch里的所有头放到一起，上身放在一起，下身放在一起. (马氏距离)
之后，根据 $S_1^{头}$ 到 $S_n^{头}$ 去select，可以根据distance去select. 同样，根据上身去下身select.
为什么要根据头→上身，上身→下身的顺序去select → 有语义上的连续性.

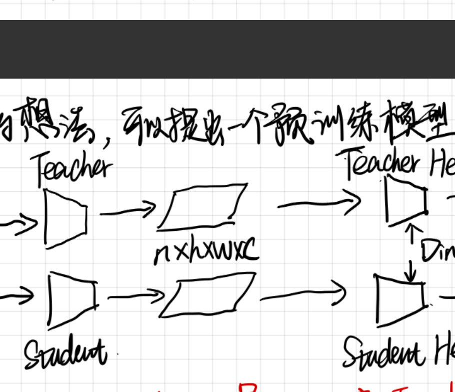
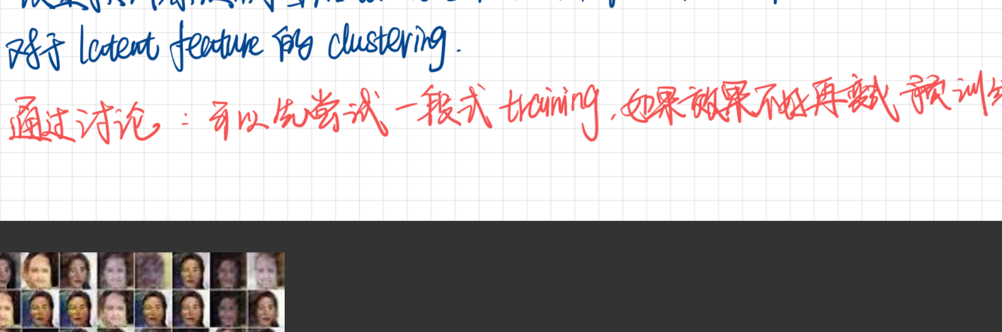


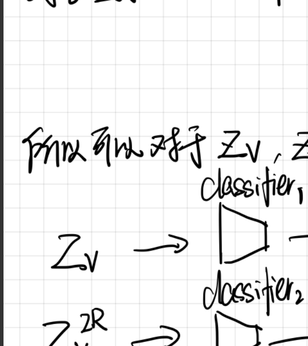
Fig. 3. Visualization of within-part inconsistency. T. Left: T is equally partitioned to $p = 6$ horizontal stripes (parts) during training. Right: Every column vector in T is denoted with a small rectangle and painted in the color of its closest part.

之后通过对比reconstruct的 $\{S_{头}, S_{上}, S_{下}\}$ 是否来自于同一个ID → Set loss
优化Encoder来extract ID-Discriminative Features.
所以PBC的作用是对person进行切分，通过match loss来优化Encoder.

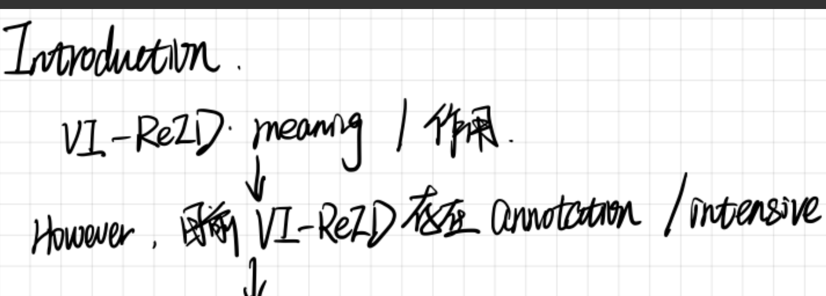
基于上面的想法，可以提出一个预训练模型



Dino Cross-entropy loss 是为了约束Teacher Encoder对于每个人都extract相似的Features. 这样对于后面的match loss也会很小.
设置预训练是由于当match loss很小时，可以不再使用PBC module，直接进行对于latent feature的clustering.
通过讨论：可以先尝试一段式training，如果效果不好再尝试预训练


If Dino Cross-entropy loss is not added, such a result may be obtained

For extracting modality-invariant features
为了基于Batch训练 → 没有必要去对所有的data进行cluster然后再进行Discrepancy loss去更新Encoder和bi-translation model
所以可以对于 Z_V, Z_V^{2R} 设置classifiers，然后通过adversarial learning 减小模态gap.



Introduction Logic Flowchart：

Introduction.
VI-ReID meaning / 作用.
↓
However，目前VI-ReID存在annotation / intensive labor
↓
所以，无监督method引起了广泛关注.
↓
引出H2H [] 和 OTLA [] 是怎么实现无监督的.
↓
但是，他们各自存在什么问题(H2H的Alignment)
OTLA用的是UDA，annotate RGB)
↓
他们都没考虑，ID-Discriminative feature 和 reasonable pull modality
↓
所以，我们提出疑问：我们是否可以在没有label的前提下让Encoder学习到ID-Discriminative feature和modality-invariant feature来减小cross-modality discrepancy.

Our contribution:
1. ID-Discriminative feature: PBC module + Mahalanobis distance + dino cross-entropy loss.
↓
优化intra-modality clustering.
2. modality-invariant feature: Bi-directional Translation + classifier loss + Adversarial learning + CMA loss.

Expressions worth learning:

- large modality gap between RGB and infrared images destroys the data manifold such that the distance of the positive RGB-IR image pairs could be greater than that of the negative RGB-RGB (or IR-IR) image pairs.
- Due to the significant difference in sensing processes, visible-infrared heterogeneous images have large appearance variations. Therefore, it's very different from conventional visible ReID problem