From last week：

- The core question addressed in this paper (what is our motivation) needs to be reorganised, clarifying the objectives of the experiments we want to compare (supervised or unsupervised, noisy labels or pseudo-labels alignment)
- Rewrite the abstract so that it is more logical, with common words or pronouns linking the upper and lower sentences to ensure that there is no break in the logic.
- Try possibilities for the methodology section.

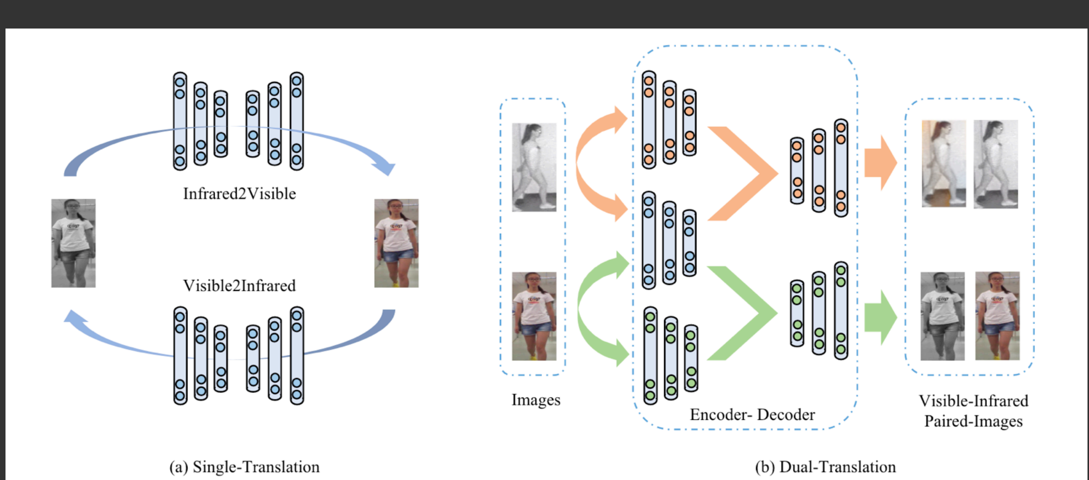
Completed in this week：

- Identified related papers for the main comparison of papers as being in the Unsupervised VI Re-ID direction and found related/similar papers for comparison.
- Found cycle GAN methods that could be applied before clustering
- Explored the value of different clustering methods to exploit for this clustering
- Explored Codebooks that can be used for clustering and alignment

---

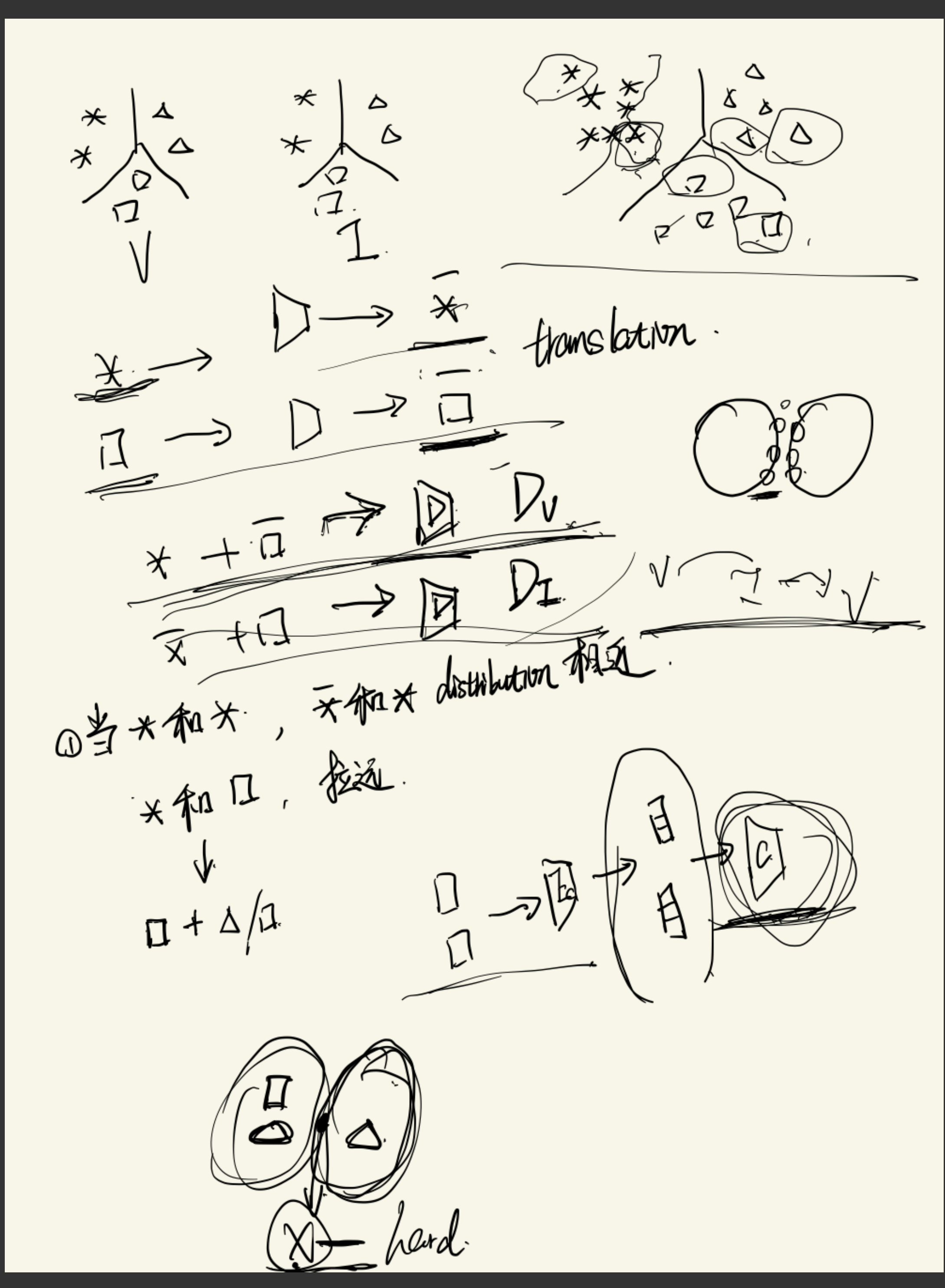
By reading the survey： Visible-Infrared Person Re-Identification: A Comprehensive Survey and a New Setting
mdpi.com/2079-9292/11/3/454

Question: Is it possible to use modality translation? --employing GAN to generate fake images destroy the structure information of generated images and introduce plenty of noise.
Therefore, the two previously proposed GANs to generate hard negative samples have many drawbacks



(a) Single-Translation (b) Dual-Translation

The sketch below shows that if image modality translation is used, when faced with edge samples of similar classes within each modality, or other hard negative samples, we can use the modality translation to determine which class the transformed image is in, and also use the cycle GAN to control that the resulting image satisfies the original distribution.



For clustering, I have an idea: first find the farthest N points (N is determined by the desire to divide into N classes), I first ensure that these N points can be connected to contain each sample, and then proceed to clustering by following the DBSCAN strategy, setting the radius of the expansion to gradually decrease during clustering. When two search circles intersect, the samples can be judged by image transformation (or set aggression games), which also allows for the exchange of information within the modalities, but requires that the distribution between the two modalities be similar, so as to facilitate the alignment - the solution: set the loss between the Set loss/lipschitz continuity between Discriminators

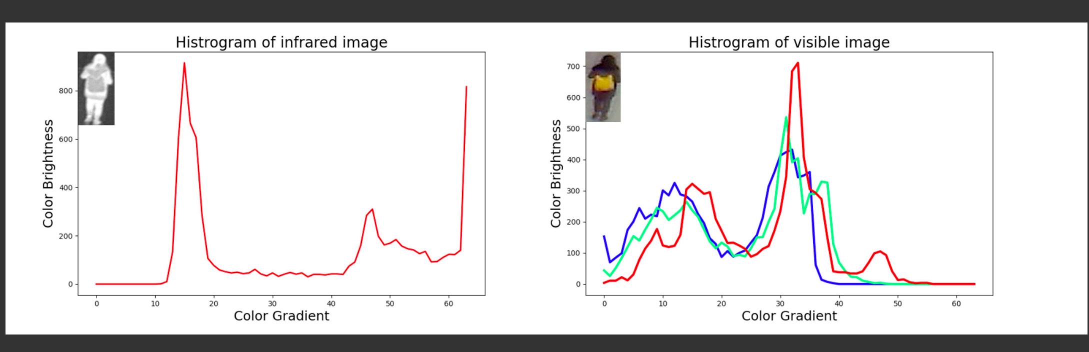
- [x] The Abstract has been rewritten in terms of the re-definition of the research question

There are very few clustering methods that suitable for two modal images — These works can hardly handle the cross-modality clustering problems because the 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.

The problem of clustering in VI Re-ID：

1. the camera view (denoted as v-factor)： intrinsic characteristic of the conventional re-ID, including significant changes in illumination, background clutter, pose, and viewpoint.
2. modality discrepancy (denoted as z-factor)： the z-factor 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,

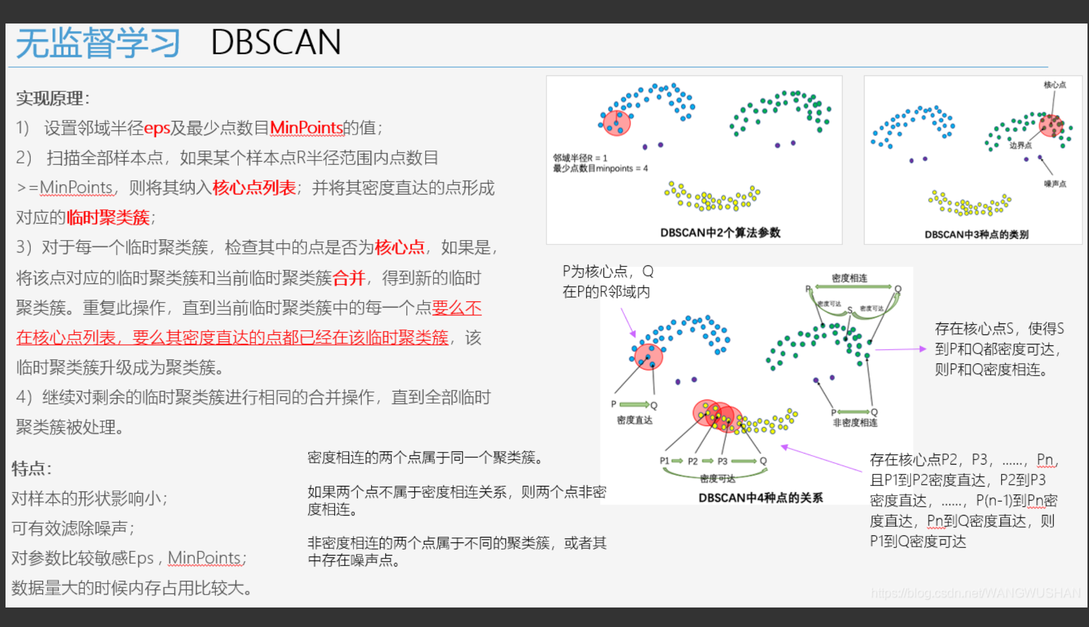
As the inter-modality discrepancy is substantially greater than the intramodality discrepancy, bridging the modality gap between the two heterogeneous modalities is a major aspect of VI-ReID research.



So we plan to use Lipschitz Continuity to reduce cross-modal discrepancy.

Question asked: What is considered a hard negative sample in cross-modal match?

Wouldn't spectral clustering and DBSCAN be better suited than K-means - for visible infrared data, which is usually characterised by low resolution and low signal-to-noise ratio?
K-means is sensitive to noise and isolated point data



DBSCAN can solve the problem of the same identity being split into two clusters in VI Re-ID, but it does not work well for isolated points.

for Spectral Clustering:

1. Spectral clustering can be computationally expensive, particularly for large datasets.
2. Parameter sensitivity
3. Spectral clustering is a non-convex optimization problem, which means that it can get stuck in local optima. This can lead to suboptimal cluster assignments and reduced performance.
4. Sensitivity to noise and outliers

- Bridge the modality gap by designing losses in the common representation space的缺点为may not be sufficient to eliminate potential heterogeneity of different modalities in the common space.
- ignore label relationships which are important for constructing semantic links between multimodal data.--Is this something we need to consider too!

An Embarrassingly Simple Approach to Semi-Supervised Few-Shot Learning（NeurIPS 2022）

Motivation: Many of the current methods for studying semi-supervised few-sample learning suffer from a number of problems:
1) the low correct rate of labelling unlabelled data with pseudo-labels, and incorrectly labelled samples can affect the final results;
2) There is a class imbalance in the pseudo-labels labelled on unlabelled data;

3) The method is more complex.

Methodology：

- Constructing a meta-task, using a pre-trained neural network as a feature extractor to extract image data, extracting features corresponding to the support set, query set and unlabeled data set in the meta-task, and training a classifier on the support set for subsequent image classification tasks；

首先通过卷积神经网络提取元任务中对应数据集合的特征：

$$x^{set \in \{S,Q,U\}} = F(I;\theta_r) \tag{1}$$

其中 I 为输入数据，$F(\cdot;\theta_r)$为预训练的卷积神经网络模型，其中θ_r为该模型的参数。x^{set}为 set 集合提取出来的特征，set 可取 S、Q 或 U，分别代表支持集、查询集以及无标签数据集。

接着初始化分类器$f(\cdot;\theta_c)$，其中θ_c为该分类器参数。用分类器将x^S映射到对应的概率空间：

$$p^S = f(x^S;\theta_c) \tag{2}$$

接着使用交叉熵损失进行训练，其中交叉熵损失表示如下：

$$L(f,y) = -\sum_k y_k p_k \tag{3}$$

- The negative pseudo-labels learning module negative the unlabelled image data with a high 95% correct rate and learns updates on the anti-labels with the classifier, iterating through until no negative pseudo-labels can be selected;
- The positive label learning module obtains category-balanced positive labels with a high 85% correct rate and learns updates with the classifier;
- The trained classifier is used on the query set to predict the final image classification category results.

Question: Can we use a similar method for labeling and then clustering?

DM2C: Deep Mixed-Modal Clustering （NeurIPS 2019）

In this paper, they chose a more challenging task:where each instance is represented in only one modality, which we call mixed-modal data.

Since there are no images of pairs, they chose to use CycleGan to build the mappings for unpaired data — motivate us！！

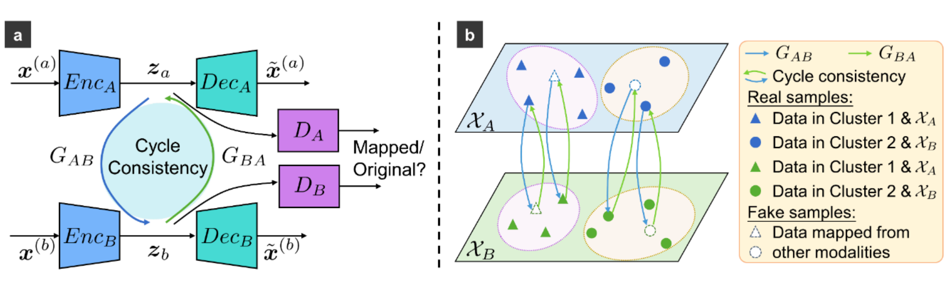


Figure 1: Overview of the proposed method. (a) Our adversarial network architecture for the unified representation learning. (b) Cycle consistency across modality-specific latent spaces illustrated on some samples. The cross-modal mappings help unify all the samples into a space.

Then a cycle-consistent mini-max game is performed on the discriminators and the mappings between modalities.--Can we learn from it?！！！

Multi-modal Alignment using Representation Codebook （CVPR 2022）

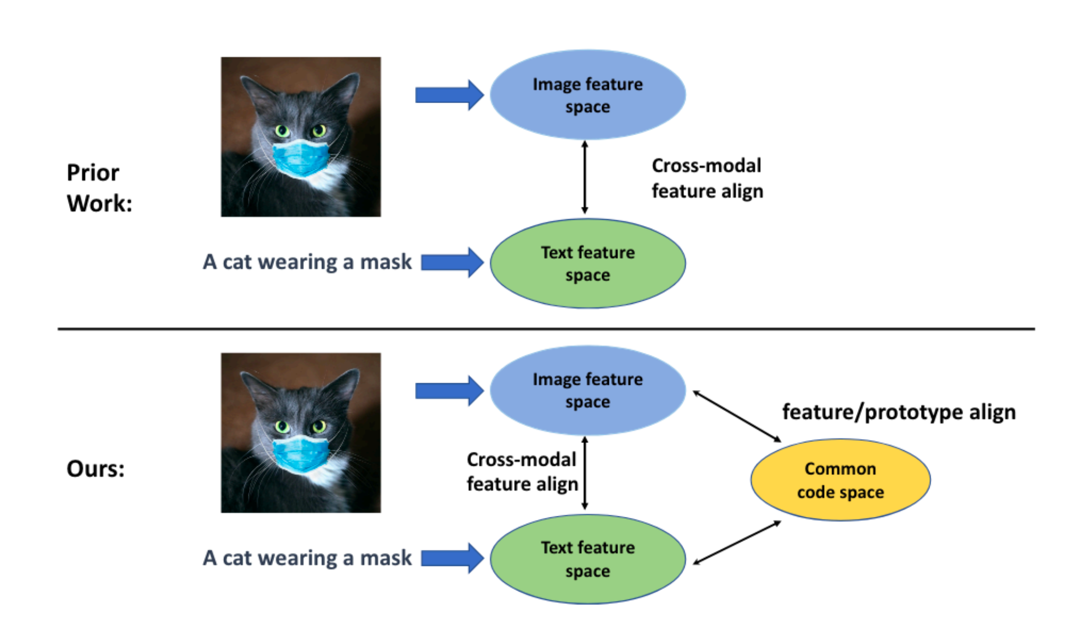


Figure 1. We propose to use a learnable codebook to better align the image and text modalities. The codebook serves as a "bridge" between the image and text features. Each codeword can be interpreted as a prototype, which enables contrasting image and text at the cluster level. We then solve an optimal transport [1] problem to optimize the distance between each modality to the prototypes, which in turn optimizes the alignment between the two modalities. Prototype vectors are learned along with the feature encoders in our V&L framework.

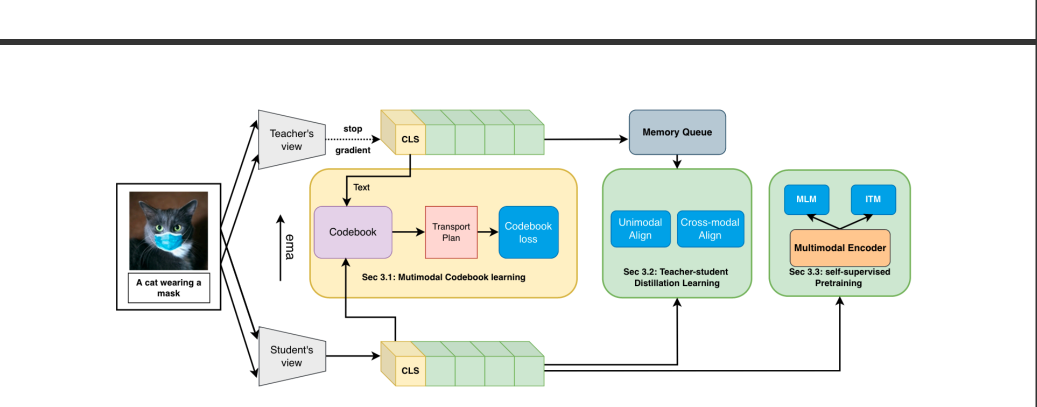


Figure 2. Overview of our framework. For simplicity, we only display a pair of teacher-student encoders (e.g., teacher for the image and student for the text) and similarly for the memory queue. The teacher is updated with an exponential moving average of the student (from the same modality). The codebook helps bridge the gap between the different modalities. The entire framework is end-to-end optimized.

A more efficient alignment strategy is proposed that uses a codebook that quantifies the space of common textual image features into codewords. These encodings or cluster centers provide a more stable means of contrastive inference than individual text or visual features. --contrastive inference across modalities (image-text) was performed - they used a learnable codebook for both image and text modalities and trained the model to predict the assignment of codewords using textual or visual information. In effect, visual and textual features are aligned during training by aligning them with a common codebook.

For a pair of images and text, can compute cross-modal similarity and intra-modal similarity as follows:-
Subsequently we can also draw on:

$$\begin{aligned}
p_{t2i}(T) &= \exp\frac{z_t z_v^{m\top}}{\gamma} \Big/ \sum_{z_v^{m'} \in \mathbf{Q}_v} \exp\frac{z_t z_v^{m'\top}}{\gamma} \\
p_{i2t}(I) &= \exp\frac{z_v z_t^{m\top}}{\gamma} \Big/ \sum_{z_t^{m'} \in \mathbf{Q}_t} \exp\frac{z_v z_t^{m'\top}}{\gamma} \\
p_{i2i}(I) &= \exp\frac{z_v z_v^{m\top}}{\gamma} \Big/ \sum_{z_v^{m'} \in \mathbf{Q}_v} \exp\frac{z_v z_v^{m'\top}}{\gamma} \\
p_{t2t}(T) &= \exp\frac{z_t z_t^{m\top}}{\gamma} \Big/ \sum_{z_t^{m'} \in \mathbf{Q}_t} \exp\frac{z_t z_t^{m'\top}}{\gamma}
\end{aligned} \tag{4}$$

where the pseudo-image negatives used to estimate pt2i(T) are drawn from the image queue Qv and used similarly for pi2t(I). -shows that strengthening one modal features facilitates cross-modal alignment--Update the similarity calculation for our VI cross-modal