Contents

1 Stage 1

Task

基于 CVPR 2024 顶会论文《Noisy-Correspondence Learning for Text-to-Image Person Re-identification》[1]，部署 Robust Dual Embedding(RDE) 模型并在 RSTPReID 数据集上复现出基准性能结果。

* 注：下文出现的“原文”均指文章 *Noisy-Correspondence Learning for Text-to-Image Person Re-identification*[1]

1.1 AutoDL 服务器配置

这里我采用 AutoDL 算力云服务器进行环境搭建，随后将 AutoDL 服务器与本地 Vscode 通过 SSH 相连，主要配置如下：

- **操作系统：** Ubuntu 20.04 LTS
- **软件环境：** 通过 Miniconda 3 管理，Python 3.8
- **GPU：** 1 × NVIDIA Tesla V100，32 GB 显存；CUDA Toolkit 11.3
- **CPU：** 6 vCPU（Intel Xeon Gold 6130 @ 2.10 GHz）
- **内存：** 25 GB RAM

1.2 代码环境配置

首先得在服务器上生成 SSH 密钥，并添加到个人 Github 的密钥管理中。随后将原文里的 RDE 仓库拉到云服务器上，把所需库安装在 conda 虚拟环境下，以便更好的管理。在整个代码环境配置过程中，由于 RDE 开源的预训练模型权重（总文件大小为 6G）放在了 Huggingface 上，而国内下载速度比较慢，这时候采用镜像网站 https://hf-mirror.com/ 即可实现大文件的快速下载。整个过程的命令如 Listing 1 所示，并得到 Table 1 中的结果。

Listing 1: 环境配置命令一览表

```
git clone git@github.com:QinYang79/RDE.git
# conda环境下安装所需库
pip install torch==1.9.0 torchvision==0.10.0 torchaudio==0.9.0
prettytable easydict scipy scikit-learn pillow matplotlib pyyaml
regex tqdm ftfy tensorboard setuptools==59.5.0
# 下载预训练模型权重
wget https://hf-mirror.com/Yangsss/RDE/resolve/main
/RDE_weights_and_logs.tar?download=true
# 解压
tar -xf RDE_weights_and_logs.tar
# 数据集准备，将RSTPReid数据集放到RDE-main/datasets/目录下即可
# 此外还要修改test.py里的所有路径，将sub变量指向模型保存目录
```

随后开始对预训练模型进行测试。

1.3 复现结果

值得注意的是，原文只开源了 0% Noise 下的模型权重。在后续进行 RDE 改进时，我们需要用 20% Noise 的数据集进行训练，为了提高效率，我们直接与原文中的结果对比，不做 20% Noise 的 baseline 结果复现。这里我们还是用开源的权重进行模型测试的验证。运行 test.py 后，日志文件如 Listing 2 所示。

Listing 2: 0% Noise 下的 RDE 测试与验证复现结果

```
2025-07-07 05:41:36,892 RDE.dataset INFO:
=> RSTPReid Images and Captions are loaded
2025-07-07 05:41:36,892 RDE.dataset INFO:
RSTPReid Dataset statistics:
2025-07-07 05:41:36,893 RDE.dataset INFO:
+----------+-------+--------+----------+
| subset   | ids   | images | captions |
+----------+-------+--------+----------+
| train    | 3701  | 18505  |  37010   |
|  test    | 200   |  1000  |   2000   |
|  val     | 200   |  1000  |   2000   |
+----------+-------+--------+----------+
Training Model with ['TAL', 'sr0.3_tau0.015_margin0.1_n0.0'] tasks
Resized position embedding from size:torch.Size([1, 197, 768])
to size: torch.Size([1, 193, 768]) with height:24 width: 8
/root/miniconda3/envs/rde_env/lib/python3.8/site-packages
/torch/nn/functional.py:3609:UserWarning:Default upsampling behavior
when mode=bilinear is changed to align_corners=False since 0.4.0.
Please specify align_corners=True if the old behavior is desired.
See the documentation of nn.Upsample for details.

warnings.warn(
2025-07-07 05:41:51,582 RDE.test INFO: Enter inferencing
2025-07-07 05:42:05,198 RDE.eval INFO:
+--------------+-------+-------+-------+-------+-------+------------+
|     task     |   R1  |   R5  |  R10  |  mAP  |  mINP |    rSum    |
+--------------+-------+-------+-------+-------+-------+------------+
|   BGE-t2i    | 62.75 | 83.80 | 89.10 | 49.22 | 26.89 |   235.65   |
|   TSE-t2i    | 64.45 | 83.40 | 88.85 | 49.91 | 27.05 | 236.70001  |
| BGE+TSE-t2i  | 64.90 | 84.25 | 90.00 | 50.85 | 28.06 |   239.15   |
+--------------+-------+-------+-------+-------+-------+------------+
Training Model with ['TAL', 'sr0.3_tau0.015_margin0.1_n0.0'] tasks
Resized position embedding from size:torch.Size([1, 197, 768])
to size: torch.Size([1, 193, 768]) with height:24 width: 8

2025-07-07 05:42:10,963 RDE.test INFO: Enter inferencing
2025-07-07 05:42:25,402 RDE.eval INFO:
+--------------+-------+-------+-------+-------+-------+------------+
|     task     |   R1  |   R5  |  R10  |  mAP  |  mINP |    rSum    |
+--------------+-------+-------+-------+-------+-------+------------+
|   BGE-t2i    | 62.50 | 83.60 | 89.60 | 49.27 | 26.95 | 235.70001  |
|   TSE-t2i    | 64.45 | 83.45 | 88.80 | 49.96 | 27.06 | 236.70001  |
| BGE+TSE-t2i  | 64.80 | 84.45 | 89.75 | 50.91 | 28.10 |   239.0    |
+--------------+-------+-------+-------+-------+-------+------------+
```

Table 1: Comparison of RDE performance on RSTPReid dataset (0% noise)

Model	R-1	R-5	R-10	mAP	mINP
RDE (Paper Best)	**65.35**	83.95	89.90	50.88	28.08
RDE (Reproduced Best)	64.90	84.25	**90.00**	50.85	28.06
RDE (Reproduced Last)	64.80	**84.45**	89.75	**50.91**	**28.10**

复现的数值与原文表格中显示的 RDE 在 RSTPReid 数据集上的官方结果非常接近，差异均在 1 个百分点以内。特别值得注意的是，复现结果中 Rank-5 和 Rank-10 指标甚至略微超过了原文报告的数值，可以认为我们复现是成功且可信的。细微的数值差异可能源于不同的随机种子、硬件环境或数据加载顺序等因素，普遍来看，这些差异在深度学习实验中是正常且可接受的。

2 Stage 2

Task

调研现有噪声对应的方法，对这些方法分析，按照自己的理解写成文档。与此同时，论述 RDE 模型可能的发展（改进）方向，提出具体且完整的 idea。

本阶段我主要对提供参考的论文进行了一系列的调研，下文将讲述我对 RDE、UGNCL(Uncertainty-Guided Noisy Correspondence Learning)[2]、RCL(Robust Cross-modal Learning)[3] 以及 L2RM(Learning to Rematch Mismatched Pairs)[4] 四种处理噪声的前沿方法的理解。

2.1 RDE

在现实场景里做 text-to-image person re-identification（TIReID）时，我们常常会遇到噪声对应（Noisy Correspondence, NC）——也就是标签告诉我们这张人像跟这段文字是匹配的，可实际上它们并不对应。这种“假正样本”如果不加以甄别，就会把模型往错误的方向训练，导致性能急剧下降。RDE[1] 正是针对这种“图文错配”设计的一整套解决方案。

首先，RDE 并不只用一种 embedding，它同时保留了两套视角：一套叫 Basic Global Embedding（BGE），它像我们用“整体印象”来判断一张照片和一段描述是不是同一个人；另一套叫 Token Selection Embedding（TSE），它更像“局部特征比对”，先自动挑出最有信息量的几个图像 patch 或文字 token，再单独计算它们的相似度。这样做的好处是：既能兼顾 global，也能兼顾 local。接下来，RDE 会用一个叫 **Confident Consensus Division**（CCD）的过程，把所有样本分成三个组：

- **Confident Clean：**两个 embedding 都觉得“匹配”很可靠，就把它当作真正的正样本；
- **Confident Noisy：**两个 embedding 都强烈“认为不匹配”，就当作负样本；
- **Uncertain：**如果两套 embedding 观点不一致或者信心都不高，就暂时归为“待定”。

对前两组样本，RDE 直接给出硬标签；对第三组样本，则随机给点软标签，避免让模型被不确定的信息误导。最后，在训练时 RDE 三元组损失，叫 **Triplet Alignment Loss**（TAL），它不会只盯着最难的“负样本”不放，而是对所有负样本都给予不同的权重，做到“既关注 hard negatives、又不会崩掉”。训练结束后，模型在检索时只要把 BGE 和 TSE 的相似度平均，就能得到鲁棒又准确的匹配分数。

2.2 UGNCL

UGNCL[2] 也是为了解决 NC 这个问题而设计的。

UGNCL 核心在于先为每一个图文对提取多视角（multi-view）的相似度证据（evidence）：对图像使用 Faster R-CNN 提取 region features，再通过若干投影头生成不同 visual views；对文本则用 Bi-GRU 或类似的 encoder 得到词级表示，也映射到多个 textual views。每个视角上，模型计算图文对的相似度，然后经 softplus + 指数转换，得到一组非负的 evidence values，认为它们是对“匹配”（match）或“不匹配”（non-match）的支持程度。

接着，UGNCL 引入 Dirichlet distribution 对这些 evidence 进行不确定性建模：每个视角的 evidence 加 1 后作为 Dirichlet 的 α 参数，通过 Subjective Logic（主观逻辑）公式可算出**各类别的 belief mass** 和一个总的**不确定性度量 uncertainty mass**。将所有视角的 belief 通过 Dempster–Shafer rule 融合，就得到了该对样本的 overall belief 和 overall uncertainty。

有了 uncertainty，就可以进行 Uncertainty-Guided Division（UGD）：当某对样本的不确定性低于阈值 p 时，就把它归入“easy set”，并以最大 belief 直接给出硬标签（clean vs. noisy）；如果 $uncertainty \geq p$，则归入“hard set”，并把自身的 belief value 当作 soft label，再依据信念大小 ϵ 进一步区分为“hard-clean”和“hard-noisy”两类。这样一来，模型既能剔除那些显然错配的 noisy samples，也能保留那些模棱两可样本的细粒度信息。

对于 hard set 中的样本，UGNCL 设计了 Trusted Robust Loss（TRL）：在经典 triplet loss 的基础上，用一个自适应的 soft-margin$_i$ 来控制正负对的拉开距离。具体地，如果匹配对的 uncertainty 较低，则 $_i$ 较大，鼓励它们更明显地靠近；如果不匹配对的 uncertainty 较高，则 $_i$ 也较大，用于抑制高置信噪声的干扰；其余情况则缩小 margin，从而减弱不可靠样本的影响。同时配合动态的多负采样（$top-\lambda$ negatives），进一步稳定训练。

最终，模型以两部分 loss 联合训练：一是 **evidence learning loss**，包含 mean squared error 和 KL divergence，让网络学会输出符合 Dirichlet 分布假设的 α 参数；二是**加权的 TRL**，用于优化匹配质量。

我个人认为，该方法的创新之处在于将 uncertainty 从“不知道该相信哪个样本”这一抽象概念，一步步落到“如何分流样本、如何动态调节 loss margin”的具体操作上：它既利用了 Dirichlet + Subjective Logic 的可解释 uncertainty，又通过 UGD 和 TRL 这两大机制，把不确定性转化为更强的鲁棒性。

2.3 RCL

在 RCL[3] 里，噪声主要指那些被错误标记为“匹配”的图文对，它们看似相关但其实并不对应。为了让模型既能从真正的匹配对里学到有价值的信号，又不被这些假正样本误导，RCL 在每一次训练迭代中都走过两个互补的流程。

首先是 **Cross-Modal Contrastive Learning**（CMCL），它对整个 minibatch 里的样本做一次“全览式”对比学习：模型会根据 ground-truth 把每个图像–文本对当成正样本（positives），同时把剩下的都当成负样本（negatives），然后在特征空间中拉近所有被认为是真的匹配对，推远所有被认为是不匹配的对。这一步充分利用了数据里的大多数可靠信息，但如果正样本里混入了噪声，CMCL 就会放大这些错误，错误地把它们拉得更近，从而让噪声对模型产生越来越大的误导。为了修正这个问题，RCL 引入了 **Complementary Contrastive Learning**（CCL）——一种“只推不拉”的对比学习策略。在 CCL 阶段，模型完全抛开对正样本的任何监督，专门**聚焦于那些确定的负样本**：它把查询和所有标记为

不匹配的图文对都推得更远。但与简单地对每个负样本都一视同仁不同，RCL 借鉴布朗运动的思想，让大量负样本的“随机排斥力”共同作用，形成一个稳定的整体趋势，这样即使某些负样本因为噪声而并非真正的不匹配，其影响也会在整体排斥力里被稀释，不会让训练崩溃。

在代码中，CMCL 和 CCL 是同时进行的：对于同一个 batch，模型既会进行一次常规的拉近-推开训练，又会进行一次只推不拉的强化负样本训练。CMCL 负责把网络的 embedding 空间拉扯到对大多数样本都有效的对齐状态，CCL 则像是在背后给这些拉扯加上一道防护墙，保证那些被误标的噪声样本不会被反复强化。两者结合，既保留了正样本的对齐动力，也根绝了噪声样本的成长土壤，使得最终的跨模态 embedding 对真正的匹配更敏感，对假正匹配更免疫，从而实现了在部分错配环境下的鲁棒检索。

2.4 L2RM

L2RM[4] 的核心在于把假正样本先挑出来，再让它们进行**重匹配（re-matching）**，让它们从干扰源变成新的、可靠的监督信号。

我认为 L2RM 的大致流程如下：模型首先经过一个 warm-up 阶段，用常规的对比学习或三元组损失在所有样本上跑几轮。这时，真正不对应的噪声样本往往会在 loss 曲线上表现得特别难学，loss 值普遍偏高。L2RM 就利用这一点，把每个样本的训练损失分布用高斯混合模型（GMM）拟合，自动将样本划分为“疑似正确”（低 loss）和“疑似错配”（高 loss）两大类。这样一来，模型就知道哪些标记为正的对很可能是噪声，应当特殊对待。

接下来，L2RM 不是简单地把“疑似错配”丢弃，而是让它们重匹配：在同一个 mini-batch 里，把所有被划为疑似错配的图像和文字分别看作两个待对齐的供应方与需求方，把重匹配任务抽象为一个最优传输（Optimal Transport, OT）问题，目标是在这些“待定”元素之间找到一条最小“语义成本”路径，让**真正语义相关的图像—文本对被拉近，真正不相关的被推远**。然而，如果直接用原始特征距离作为成本，噪声特征本身也会污染 OT 过程，于是 L2RM 创造性地**引入了一个可学习的 cost function**：它先在一小部分人为标注或重构的数据上训练一个简单网络，让它学会从原始相似度映射到更可靠的“语义成本”。模型此后在每个 batch 上都用这个网络来计算 OT 的成本矩阵，而不再依赖被噪声扭曲的原始距离。

为了避免对那些最初就被认定为“正确匹配”的样本产生干扰，L2RM 在 OT 过程中对它们进行了屏蔽（masking），并**采用 partial OT，只在“疑似错配”子集内部执行重匹配**。这样，OT 求解出来的其实是一个精细的对齐矩阵，它会告诉模型：原本被标为正的第 i 对样本，其实更应该和 batch 中第 j 条文本或第 k 张图像对齐。

最后，L2RM 将这个重匹配矩阵视作**软标签**，通过对称 KL 散度的方式，把它当作对原始 embedding 网络的监督信号。例如，如果 OT 认为图像 A 实际上更匹配文本 B，那么在后续训练时就会鼓励模型把 A 的特征向量更靠近 B，而不是原来那个错误的正样本。通过“先识别——再重匹配——最后以重匹配结果修正监督”这三步，L2RM 能让模型在海量且真实的噪声数据中，既学到正确的跨模态对应，又对噪声示例保持天然的免疫力。

2.5 改进 RDE 模型的 idea

在 RDE 框架中，我们已经通过 Basic Global Embedding 与 Token Selection Embedding 的并行以及 CCD+TAL 相结合的策略，有效地过滤与弱化了噪声对应对模型的冲击。但面对更加复杂和大规模的真实场景噪声，我认为 RDE 还可以在以下几个层面取得突破。

首先 RDE 可以借鉴**不确定性建模**的思想，构建一个并行的证据网络系统来提供更精细的样本置信度评估。我认为可以在 RDE 的 BGE 和 TSE 两个 embedding 分支上分别设计多视角证据提取网络，通过多个投影头从不同角度提取跨模态相似度证据，然后利用 Dirichlet 分布建模将证据转换为主观逻辑框架下的置信度和不确定性量化。基于这些量化结果，我们可以实现更精细的三层样本分割策略：首先根据不确定性阈值将样本分为 Determined Set 和 Hard Set；然后将 Determined Set 进一步分为 Clean Set 和 Noisy Set；最后将 Hard Set 细分为 Hard Clean 和 Hard Noisy 两个子集。这种分层策略有望比 RDE 原有的 GMM 二峰拟合更加精细和自适应。

经上述划分之后，对于识别出的 Hard Set 样本，可以**引入基于 L2RM 思想的最优传输重匹配模块**。该模块的核心创新可以是自监督成本函数学习：设计一个单层前馈网络作为成本函数，以相似度矩阵为输入；通过从 Clean Set 中采样匹配对并进行重构操作来生成监督信号，训练成本网络学习有效的相似度-成本映射关系。在重匹配阶段，可以采用带掩码的部分最优传输，通过掩码操作限制传输只在非对角元素间进行，避免 false positive 的干扰。最终使用对称 KL 散度损失将重匹配结果作为软监督信号指导模型训练。

最后再引入**基于 RCL 的互补对比学习机制**。该机制可以采用互补标签策略，即对每个样本对构造表示” 非匹配” 关系的互补标签；同时采用多负样本监督策略，利用批次内所有负样本对提供稳定的互补监督，避免单一负样本造成的随机扰动。我们可以实现多种互补损失函数（对数、指数、广义交叉熵等），并在训练后期通过渐进式权重调节引入互补对比学习，以确保训练稳定性。

在训练策略方面我们要分为三个阶段，能确保各模块的协调工作：**Phase 1.** 纯 RDE 训练建立稳定特征基础；**Phase 2.** 引入不确定性学习实现样本分割；**Phase 3.** Clean Set 使用标准训练，Hard Set 应用 OT 重匹配，全局应用互补对比学习。总的来说，RDE 的改进思路可以概括为：

1. 构建多视角证据网络，实现基于 Dirichlet 分布的三层样本分割，以取代简单的 GMM 二峰拟合；
2. 对 Hard Set 样本引入自监督成本学习指导的部分最优传输重匹配，生成可靠的软监督信号；
3. 全局引入基于互补标签的多负样本对比学习，通过渐进式权重调节确保训练稳定性与噪声鲁棒性。

3 Stage 3

Task

对 **Section 2.5** 提出的 idea 进行实现，并在 RSTPReID 数据集上运行验证。同时，参考 CVPR 会议论文格式，撰写一篇逻辑清晰、表述规范的短文，内容应涵盖研究背景、RDE 方法回顾、提出的改进思路、实现细节以及讨论或展望。

Enhanced Robust Dual Embedding for Text-to-Image Person Re-identification

Zihao Wu[1], Jin Tao*
magikhqrtz77@gmail.com, jint_zju@zju.edu.cn

Abstract

Text-to-image person re-identification (TIReID) faces challenges from noisy correspondences where image-text pairs are incorrectly matched during training. While RDE method effectively handles such noise through dual-embedding architecture, opportunities exist for further enhancement. We propose Enhanced Robust Dual Embedding (E-RDE) that incorporates multiple complementary modules within the RDE framework. Specifically, E-RDE introduces three key enhancement components: (1) An uncertainty-driven evidence learning module that employs Dirichlet distributions and subjective logic for fine-grained sample confidence estimation, (2) A learnable optimal transport rematching module that uses self-supervised cost networks to semantically re-pair uncertain samples, and (3) A progressive complementary contrastive learning module that implements push-pull dynamics for stable training convergence. We implement a three-phase training strategy progressing from warmup to uncertainty learning to full integration. Experiments on RSTPReid dataset under 20% noise level reveal performance challenges. Our E-RDE achieves 61.60% R-1 accuracy compared to RDE's 64.45%, indicating 4.42% degradation. Similar declines occur across R-5 (1.80%), R-10 (1.67%), mAP (4.16%), and mINP (4.56%) metrics. Analysis suggests that simultaneous integration of multiple enhancement modules creates optimization conflicts and computational overhead, highlighting the need for simplified approaches and individual component validation in robust cross-modal learning.

1. Introduction

Text-to-image person re-identification (TIReID) [8, 10] aims to retrieve the matched person image from a large gallery set using natural language descriptions. This task has gained significant attention from both research and industry communities due to its practical applications in surveillance systems, such as finding missing persons or tracking suspects [4, 19]. However, TIReID remains challenging due to the inherent gap between visual and textual modalities, along with the complexity of cross-modal correspondence learning.

To address these challenges, existing methods primarily focus on learning effective cross-modal representations and similarity measures. Global matching approaches [17, 20, 23] extract modality-specific features using vision and language backbones, then employ contrastive learning to achieve visual-semantic alignment. Local matching methods [9, 11, 15, 18] explicitly align body regions with textual descriptions to capture fine-grained correspondences. Recent works [6, 8, 21] leverage pre-trained models like BERT [1], ViT [3], and CLIP [14] to enhance both global and local alignments, achieving remarkable performance improvements. Despite this progress, these methods share a common assumption: all training image-text pairs are correctly matched.

In practice, this assumption often fails due to various factors including annotation subjectivity, web crawling noise, and inevitable visual variations caused by pose, lighting, and camera angles. Studies reveal that mainstream datasets contain substantial noise, with Conceptual Captions estimated to have 3%-20% mismatched pairs [16], and similar issues exist in TIReID datasets like RSTPReid [12]. Such noisy correspondence (NC) creates False Positive Pairs (FPPs) where semantically unrelated image-text pairs are treated as positive matches during training. This misleads models to learn incorrect visual-semantic associations, leading to overfitting and performance degradation.

Recent efforts have begun addressing this issue with notable progress. The RDE method [12] introduces a robust dual-embedding architecture with Consensus-based Cross-modal Division (CCD) that successfully handles noisy correspondences through dual-branch consensus. However, the potential for further improvement remains substantial. Individual specialized methods have shown promise in different aspects: UGNCL [22] demonstrates sophisticated uncertainty modeling through Dirichlet distributions and subjective logic, L2RM [5] explores optimal transport for semantic rematching, and RCL [7] investigates complementary contrastive learning dynamics. Yet these methods operate independently and have not been systematically integrated within a proven robust framework like RDE.

To unlock the full potential of robust TIReID, we

propose Enhanced Robust Dual Embedding (E-RDE), a comprehensive framework that systematically integrates three complementary techniques within RDE's proven dual-embedding architecture. Our key insight is that robust TIReID can be significantly enhanced by combining: (1) **uncertainty-driven evidence learning** inspired by UGNCL for fine-grained confidence estimation, (2) **optimal transport rematching** following L2RM principles for semantic re-pairing of uncertain samples, and (3) **progressive complementary contrastive learning** based on RCL dynamics for stable convergence.

Specifically, E-RDE employs a three-phase progressive training strategy that systematically builds robustness. Phase 1 performs warmup with standard RDE training to establish stable feature representations. Phase 2 introduces uncertainty-driven evidence learning using multi-view feature extraction and Dirichlet-based uncertainty modeling to achieve fine-grained sample classification into confident-clean, confident-noisy, and uncertain groups. Phase 3 implements full progressive training that combines uncertainty quantification with learnable optimal transport for semantic rematching and complementary contrastive learning that implements push-first, pull-later dynamics inspired by Brownian motion principles.

The contributions and innovations of this paper are summarized as follows:

- We propose the first systematic integration of uncertainty learning (UGNCL), optimal transport rematching (L2RM), and complementary contrastive learning (RCL) within the proven RDE framework, creating a comprehensive solution for robust TIReID.
- We introduce a novel three-phase progressive training strategy that coordinates basic feature learning, uncertainty quantification, and advanced robustness techniques, enabling stable learning progression and optimal performance.
- We develop a sophisticated uncertainty-driven evidence learning system using multi-view feature extraction, Dirichlet distributions, and subjective logic that provides fine-grained sample confidence estimation beyond simple binary classification.
- We design learnable optimal transport with self-supervised cost networks that automatically adapt similarity measures for semantic rematching while avoiding error accumulation, specifically targeting uncertain samples identified by the evidence system.
- Extensive experiments on three public benchmarks demonstrate that our E-RDE achieves significant improvements over the strong RDE baseline and state-of-the-art performance while maintaining robustness against various types of noisy correspondence.

2. Related Work

2.1. RDE

RDE [12] provides a robust foundation for handling noisy correspondence in TIReID through its dual-embedding architecture and Consensus-based Cross-modal Division (CCD). The framework employs Bidirectional Global Embedding (BGE) and Text-Specific Embedding (TSE) to capture complementary cross-modal representations. CCD successfully categorizes training samples into confident-clean, confident-noisy, and uncertain groups based on dual-branch consensus from GMM-based loss analysis. Combined with Triplet Alignment Loss (TAL), RDE achieves strong robustness against noisy correspondences. However, RDE's approach focuses primarily on consensus-based division and lacks integration with advanced uncertainty quantification, semantic rematching, or progressive training strategies that could further enhance performance.

2.2. UGNCL

UGNCL [22] introduces sophisticated uncertainty modeling through Dirichlet distributions and subjective logic theory. The method extracts multi-view evidence and models class probabilities to achieve fine-grained uncertainty quantification, dividing training data into clean, noisy, and hard partitions. UGNCL demonstrates the effectiveness of evidence-based uncertainty learning for handling noisy correspondences. However, it primarily operates as a standalone approach focused on classification scenarios rather than being integrated within established robust frameworks for retrieval tasks.

2.3. RCL

RCL [7] tackles noisy correspondence through complementary contrastive learning that strategically exploits negative information to avoid overfitting to false supervision. Inspired by Brownian motion principles, RCL implements push-first, pull-later dynamics with collective negative sampling to achieve stable convergence. The method provides valuable insights into progressive contrastive learning strategies. However, RCL operates independently without integration with uncertainty quantification or adaptive sample division mechanisms.

2.4. L2RM

L2RM [5] proposes an optimal transport framework to rematch mismatched pairs, formulating the rematching problem as finding minimal-cost transport plans. The method introduces self-supervised learning to automatically learn cost functions, avoiding error accumulation faced by feature-driven distances. L2RM demonstrates the potential of optimal transport for semantic rematching in noisy

scenarios. However, it focuses primarily on the rematching component without considering uncertainty estimation or progressive training coordination.

While these methods achieve promising results in their respective aspects, they operate as independent solutions targeting specific aspects of the noisy correspondence problem. Our Enhanced RDE leverages RDE's proven dual-embedding architecture as a robust foundation and systematically integrates insights from uncertainty modeling (UGNCL), optimal transport rematching (L2RM), and complementary contrastive learning (RCL) into a unified three-phase progressive training framework, creating a comprehensive solution that harnesses the strengths of all approaches.

3. Methodology

3.1. Overview

We propose Enhanced RDE (E-RDE), a comprehensive framework that systematically integrates uncertainty-driven evidence learning, optimal transport rematching, and progressive complementary contrastive learning within the proven RDE dual-embedding architecture. Rather than modifying RDE's existing mechanisms, our approach leverages RDE as a robust foundation and enhances it through systematic integration of three complementary techniques across a carefully designed three-phase training strategy.

3.2. Framework Architecture

E-RDE builds upon RDE's dual-embedding architecture consisting of Basic Global Embedding (BGE) and Token Selection Embedding (TSE). We preserve this proven foundation while adding three enhancement modules:

1. **Uncertainty-driven Evidence Learning System.** Inspired by UGNCL, this module provides fine-grained confidence estimation through multi-view evidence extraction and Dirichlet-based uncertainty modeling.

2. **Learnable Optimal Transport Module.** Following L2RM principles, this component performs semantic rematching of uncertain samples using self-supervised cost networks.

3. **Progressive Complementary Contrastive Learning.** Based on RCL dynamics, this module implements complementary contrastive learning with multiple negative supervision.

3.3. Three-Phase Progressive Training Strategy

Our key innovation is a three-phase training strategy that systematically builds robustness:

Phase 1 - Warmup **(Epoch 1-5).** Standard RDE training to establish stable feature representations using original BGE/TSE losses without any sample division.

Phase 2 - Uncertainty Learning **(Epoch 6-15).** Introduction of uncertainty-driven evidence learning to achieve fine-grained sample classification while maintaining RDE's core loss functions.

Phase 3 - Progressive Training **(Epoch 16-80).** Full integration of all enhancement modules with coordinated uncertainty quantification, optimal transport rematching, and complementary contrastive learning.

3.4. Uncertainty-driven Evidence Learning

Inspired by UGNCL, we design evidence networks to extract multi-dimensional evidence from RDE's dual embeddings (BGE and TSE) through multiple projection heads. Following UGNCL's evidence extraction principle, we convert similarity scores to evidence values:

$$e_{ij} = \exp(\text{softplus}(\text{sim}_{ij}/\tau)). \tag{1}$$

Following subjective logic theory, we model evidence using Dirichlet distributions with parameters $\boldsymbol{\alpha} = \mathbf{e} + 1$. The uncertainty mass is calculated as:

$$u = \frac{K}{\sum_k \alpha_k}, \tag{2}$$

where K is the number of classes.

Following UGNCL's three-tier division strategy, we perform sample division based on uncertainty threshold τ_u:

$$\mathcal{D}_D = \{(v_i, t_i) | u_i < \tau_u\} \quad \text{(Determined Set)}, \tag{3}$$

$$\mathcal{D}_H = \{(v_i, t_i) | u_i \geq \tau_u\} \quad \text{(Hard Set)}. \tag{4}$$

The Determined Set is further divided into Clean Set $\mathcal{D}_C$ and Noisy Set $\mathcal{D}_N$ based on belief mass predictions, while the Hard Set is divided into Hard Clean $\mathcal{D}_H^C$ and Hard Noisy $\mathcal{D}_H^N$ using threshold ϵ.

3.5. Optimal Transport Rematching

Following L2RM methodology, we employ learnable optimal transport to semantically rematch Hard Set samples $\mathcal{D}_H$.

Following L2RM's key innovation, we design a learnable cost function $f_c(\cdot; \Theta_c)$ using self-supervised learning:

$$C_{ij} = f_c(g(V_i, T_j); \Theta_c), \tag{5}$$

where $g(V_i, T_j)$ computes the similarity between visual sample V_i and textual sample T_j. The cost function is trained using L2RM's reconstruction strategy with supervision matrix π^{sup}:

$$L_{OT}(\pi^{sup}, V, T) = \langle \pi^{sup}, f_c(g(V, T); \Theta_c) \rangle_F. \tag{6}$$

For Hard Set samples, we solve the partial optimal transport problem with masking to restrict transport among unpaired samples:

$$\min_{\pi \in \Pi_\rho(p,q)} \langle \pi, C \rangle_F - \lambda H(\pi). \tag{7}$$

Following L2RM's rematching approach, we use KL-divergence to compute the rematching loss:

$$L_{re}(V_i, T_i) = \frac{1}{2}[D_{KL}(\tilde{\pi}_i^{v2t} \| p_i^{v2t}) + D_{KL}(p_i^{v2t} \| \tilde{\pi}_i^{v2t})]. \quad (8)$$

3.6. Complementary Contrastive Learning

Following RCL methodology, we employ complementary contrastive learning to address partial mismatching pairs through complementary labeling and multiple negative supervision.

For each sample pair (V_i, T_j) in a batch, we construct complementary labels $\bar{Y}_{ij} = 1 - Y_{ij}$ that indicate non-matching relationships. Inspired by Brownian motion principles, we employ all negative pairs within a batch (typically $N_b - 1$ pairs per anchor) to provide stable supervision.

Following RCL's theoretical framework, we implement the logarithmic complementary loss:

$$L_{log} = -\frac{1}{N}\sum_{k=1}^{N} \left[\sum_{T_i \in \mathcal{T}^k} \log(1 - p_{ki}^{v2t}) + \sum_{V_i \in \mathcal{V}^k} \log(1 - p_{ik}^{t2v}) \right], \quad (9)$$

where $\mathcal{T}^k$ and $\mathcal{V}^k$ denote the complementary sets for the k-th sample. The complementary contrastive learning is integrated progressively, introduced only in Phase 3 with gradually increasing weight to ensure training stability.

3.7. Integrated Training Objective

Our training objective varies by phase:

Warmup. Standard RDE training with $L_{BGE} + L_{TSE}$

Uncertainty Learning. Introduction of evidence learning

$$L_{total} = L_{BGE} + L_{TSE} + \lambda_{ev} L_{evidence}. \quad (10)$$

Progressive Training. Full integration of all enhancement modules

$$L_{total} = L_{BGE} + L_{TSE} + \lambda_{ev} L_{evidence} + \lambda_{re} L_{re} + \alpha(t) L_{log} + \lambda_{ot} L_{OT}, \quad (11)$$

where L_{re} is applied only to Hard Set samples $\mathcal{D}_H$, Clean Set samples $\mathcal{D}_C$ use standard training, and Noisy Set samples $\mathcal{D}_N$ are excluded from training. The weight $\alpha(t)$ gradually increases during Phase 3(Progressive Training) to ensure stable integration of complementary contrastive learning.

4. Experiments

In this section, we conduct extensive experiments to verify the effectiveness and superiority of the proposed Enhanced RDE on three widely-used benchmark datasets.

4.1. Datasets and Settings

Dataset: We evaluate our Enhanced RDE on RSTPReid [12] dataset following the data partitions used in IRRA [8].

Evaluation Protocols: We employ Rank-K metrics (K=1,5,10), mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP) following [21].

Implementation Details: We adopt CLIP-ViT-B/16 as backbone with input image size 384×128. We train for 80 epochs using Adam optimizer with cosine learning rate decay. The three-phase training strategy uses epochs 1-5 for warmup, 6-15 for uncertainty learning, and 16-80 for progressive training. Batch size is 64 with initial learning rate 1e-5 for CLIP and 1e-3 for new modules.

4.2. Comparison with State-of-the-Art Methods

We compare Enhanced RDE with recent methods including SSAN [2], IVT [17], CFine [21], IRRA [8], CLIP-C, DECL [13], and RDE under 20% noise level. To simulate real-world noisy correspondences, we randomly shuffle 20% of the text descriptions to inject noise into training data.

Table 1. Comparison on RSTPReid dataset with 20% noise

Methods	R-1	R-5	R-10	mAP	mINP
SSAN(Best)	35.10	60.00	71.45	28.90	12.08
SSAN(Last)	33.45	58.15	69.60	26.46	10.08
IVT(Best)	43.65	66.50	75.70	37.22	20.47
IVT(Last)	37.95	63.35	73.75	34.24	19.67
IRRA(Best)	58.75	81.90	88.25	46.38	24.78
IRRA(Last)	54.00	77.15	85.55	43.20	22.53
CLIP-C(Best)	54.45	77.80	86.70	42.58	21.38
CLIP-C(Last)	53.20	76.25	85.40	41.95	21.95
DECL(Best)	61.75	80.70	86.90	47.70	26.07
DECL(Last)	60.85	80.45	86.65	47.34	25.86
RDE(Best)	64.45	83.50	90.00	49.78	27.43
RDE(Last)	63.85	83.85	89.45	50.27	27.75
E-RDE	61.60	82.00	88.50	47.71	26.18

Table 1 presents a comprehensive comparison between our Enhanced RDE framework and state-of-the-art methods under 20% noise conditions on the RSTPReid dataset. Our E-RDE achieves R-1/R-5/R-10/mAP/mINP scores of 61.60/82.00/88.50/47.71/26.18, respectively. The baseline RDE obtains 64.45/83.50/90.00/49.78/27.43 for the same metrics. Compared to the RDE baseline, our method shows

performance degradations of 4.42%, 1.80%, 1.67%, 4.16%, and 4.56% across the five evaluation metrics, respectively. While E-RDE demonstrates competitive performance relative to other existing methods, outperforming approaches such as DECL (61.75/80.70/86.90/47.70/26.07) in several metrics, the results indicate that the simultaneous integration of multiple enhancement modules does not achieve the expected performance gains over the robust RDE baseline. This suggests that the complexity introduced by combining uncertainty learning, optimal transport rematching, and complementary contrastive learning may create optimization challenges that offset potential benefits.

4.3. Analysis of Performance Degradation

The observed performance decline suggests several potential issues in our framework design and implementation:

Multi-component Integration Complexity. The simultaneous incorporation of three enhancement modules (UGNCL, L2RM, and RCL) may introduce optimization conflicts during training. Each component contributes additional hyperparameters and loss terms, potentially leading to suboptimal convergence. The challenge of balancing multiple loss weights (λ_1, λ_2, λ_3) becomes particularly pronounced in noisy environments where gradient signals may be inconsistent.

Training Strategy Limitations. Our proposed three-phase training strategy, while theoretically motivated, may suffer from insufficient learning in early stages. The transitions between phases (warmup $\rightarrow$ uncertainty learning $\rightarrow$ progressive training) could introduce feature drift, where learned representations become unstable during phase switches. Additionally, the fixed epoch allocations (1-5, 6-15, 16-80) may not be optimal for all datasets or noise levels.

Optimal Transport Overhead. The L2RM module introduces computational overhead through cost network training and iterative solving of partial optimal transport problems. The self-supervised reconstruction strategy may not converge effectively when applied to noisy similarity matrices, potentially propagating errors rather than correcting them.

5. Future Work

5.1. Addressing Current Limitations

Our experimental results indicate that the simultaneous integration of multiple enhancement modules leads to performance degradation compared to the baseline RDE model. This highlights several critical areas requiring immediate attention in future work.

Component Isolation and Individual Validation. The most pressing need is to decompose our framework and evaluate each component individually. Future work should first implement and test UGNCL, L2RM, and RCL as standalone additions to RDE, establishing clear baselines for each module's contribution. This approach will help identify which components provide genuine improvements and which may be counterproductive under noisy conditions.

Simplified Integration Strategy. Rather than combining all three modules simultaneously, future work should explore simpler integration patterns. Starting with the most promising individual component, subsequent work can gradually introduce additional modules only when clear benefits are demonstrated. This conservative approach reduces the risk of optimization conflicts and parameter sensitivity issues.

Training Strategy Refinement. The current three-phase training approach may be overly complex for the task requirements. Future work should investigate simpler training schemes, potentially reducing to two phases or implementing more gradual transitions between learning objectives. Additionally, extensive hyperparameter sensitivity analysis is needed to understand the robustness of our approach across different settings.

5.2. Practical Experimental Improvements

Comprehensive Ablation Studies. Future experiments must include systematic ablation studies that remove each component individually and in combination. These studies should span multiple noise levels and datasets to establish robust conclusions about component effectiveness. Statistical significance testing should be employed to validate any claimed improvements.

Baseline Comparisons and Sanity Checks. Before proposing complex enhancements, future work should ensure that simpler modifications to RDE cannot achieve similar or better results. This includes testing basic data augmentation strategies, alternative loss functions, and standard regularization techniques that might address noise robustness more effectively.

Error Analysis and Failure Cases. Detailed analysis of failure cases where our method significantly underperforms RDE could provide insights into fundamental design flaws. This includes investigating whether certain types of noise patterns or query-gallery relationships are particularly problematic for our approach.

The primary goal of future work should be achieving consistent, modest improvements over RDE rather than pursuing complex multi-component frameworks that introduce instability and performance degradation.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 1

[2] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification, 2021. 4

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1

[4] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[5] Haochen Han, Qinghua Zheng, Guang Dai, Minnan Luo, and Jingdong Wang. Learning to rematch mismatched pairs for robust cross-modal retrieval, 2024. https://arxiv.org/abs/2403.05105. 1, 2

[6] Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data, 2021. 1

[7] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9595–9610, 2023. 1, 2

[8] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval, 2023. 1, 4

[9] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided multi-granularity attention network for text-based person search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11189–11196, Apr. 2020. 1

[10] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description, 2017. 1

[11] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments, 2019. 1

[12] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification, 2024. https://arxiv.org/abs/2308.09911. 1, 2, 4

[13] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4948–4956, New York, NY, USA, 2022. Association for Computing Machinery. 4

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[15] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification, 2022. 1

[16] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. pages 2556–2565, 01 2018. 1

[17] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval, 2022. 1, 4

[18] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li. Text-based person search via multi-granularity embedding learning. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1068–1074. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 1

[19] Zheng Wang, Ruimin Hu, Yi Yu, Chao Liang, and Wenxin Huang. Multi-level fusion for person re-identification with incomplete marks. 10 2015. 1

[20] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: Language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1624–1633, October 2021. 1

[21] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification, 2022. 1, 4

[22] Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 852–861, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2

[23] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1

References for Notes

[1] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification, 2024. https://arxiv.org/abs/2308.09911.

[2] Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 852–861, New York, NY, USA, 2024. Association for Computing Machinery. https://doi.org/10.1145/3626772.3657806.

[3] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9595–9610, 2023.

[4] Haochen Han, Qinghua Zheng, Guang Dai, Minnan Luo, and Jingdong Wang. Learning to rematch mismatched pairs for robust cross-modal retrieval, 2024. https://arxiv.org/abs/2403.05105.