UniCat: Crafting a Stronger Fusion Baseline for Multimodal Re-Identification

Jennifer Crawford
Modern Intelligence
jenni@modernintelligence.ai

Haoli Yin
Modern Intelligence
haoli@modernintelligence.ai

Luke McDermott
Modern Intelligence
luke@modernintelligence.ai

Daniel Cummings
Modern Intelligence
daniel@modernintelligence.ai

Editors: Marco Fumero, Emanuele Rodolà, Clementine Domine, Francesco Locatello, Gintare Karolina Dziugaite, Mathilde Caron

Abstract

Multimodal Re-Identification (ReID) is a popular retrieval task that aims to re-identify objects across diverse data streams, prompting many researchers to integrate multiple modalities into a unified representation. While such fusion promises a holistic view, our investigations shed light on potential pitfalls. We uncover that prevailing late-fusion techniques often produce suboptimal latent representations when compared to methods that train modalities in isolation. We argue that this effect is largely due to the inadvertent relaxation of the training objectives on individual modalities when using fusion, what others have termed modality laziness. We present a nuanced point-of-view that this relaxation can lead to certain modalities failing to fully harness available task-relevant information, and yet, offers a protective veil to noisy modalities, preventing them from overfitting to task-irrelevant data. Our findings also show that unimodal concatenation (UniCat) and other late-fusion ensembling of unimodal backbones, when paired with best-known training techniques, exceed the current state-of-the-art performance across several multimodal ReID benchmarks. By unveiling the double-edged sword of "modality laziness", we motivate future research in balancing local modality strengths with global representations.

1 Introduction

Multimodal representation learning has rapidly grown in popularity as deep learning researchers are continually innovating new ways to combine different channels of information simultaneously. Common tasks in this field revolve around using text and images for input/output [1] or even using image-paired data as a basis for learning a joint embedding across a wide variety of modalities (e.g., text, audio, IMU, depth) [3]. However, an often overlooked subfield is in the area of multimodal re-identification (ReID), where the objective is to match instances of the same object across different sensor sources and type (e.g., visible spectrum, infrared). In many cases, ReID entails learning object representations from multiple views such as in security or retail applications, where one might want to identify if an object appearing in one camera's field-of-view is the same object appearing in another camera later in time. In this setting, each modality can have its strengths and limitations. For instance, although a regular RGB image has color information, it will be affected by illumination changes (e.g., daytime vs. nighttime), whereas an infrared image might not.

Proceedings of the I edition of the Workshop on Unifying Representations in Neural Models (UniReps 2023).

Despite their ambitious goals, many recent fusion methods encounter limitations. A notable challenge is *modality laziness* as described by Du et al. [2]. In conventional late fusion techniques that employ shared loss functions between modalities—such as vanilla late fusion through concatenating or averaging representations, where the cross-entropy (CE) loss is determined after fusion [13]—there is potential for individual modalities to contribute less to the overall training classification accuracy. This concern is especially evident in downstream retrieval tasks like ReID. Based on findings by Peng et al. [9], one could infer that when modalities are trained individually, the representation of a single modality will have a direct influence on the loss function since it's the sole contributor in that setting.

Guided by these challenges, this study zeroes in on evaluating fusion techniques within the ReID context. While our lens is sharply focused on ReID, the derived insights hold broader relevance, extending to a vast spectrum of multimodal tasks that utilize late fusion techniques. Our main contributions begin with creating a comprehensive evaluation of late fusion techniques in the current multimodal ReID landscape, spotlighting their inherent limitations from a modality laziness perspective. Next, we provide empirical evidence establishing late-fusion unimodal ensembled models as a new strong baseline, outperforming existing state-of-the-art methods in multimodal ReID benchmarks. Finally, we discuss plausible explanations for this surprising performance behavior in the context of unimodal model ablations and future research directions for better understanding multimodal representation learning and model design.

2 Preliminaries and Method

Multimodal representation learning focuses on the efficient and effective compression of high-dimensional data, especially for downstream tasks like ReID. For this purpose, convolutional neural networks (e.g., ResNet-50) and Vision Transformers (ViTs) are typically employed as visual encoders [7]. Within this context, *late fusion* stands out as the predominant method for integrating multiple data modalities. This approach is favored over the less-common *early fusion*, which integrates multiple data modalities before inputting them into a single model. Notably, early fusion can be sensitive to pre-processing and lacks scalability [10]. The late fusion approach first encodes each modality separately before combining their outputs, usually via averaging or concatenation mechanisms. While late fusion does not allow for cross-modal interactions, it is one of the most widespread fusion strategies due to the intuition that the approach is able to leverage already existing and carefully optimized architectures for the distinct modalities.

In object ReID, the task is to match objects across different data captures, typically from disparate views or time instances. Multimodal ReID datasets are given by n object samples, where each sample consists of M coincidental images (corresponding to M modalities) and an identifying label (ID): for each sample we have $[(x_1, x_2, ..., x_M), y]$ where $x_i \in \mathbb{R}^{C \times H \times W}$ is the i-th modality image of the object and $y \in \mathcal{D}$ denotes its ID. We consider *multiple views* of an object to be different samples that share the same object ID. Of note, the IDs we see during train-time do not overlap with those seen at inference. As such, object ReID is often solved not by using the direct output of a classifier, but by bringing different views of the same object together in the latent space and pushing different objects away from each other. This is commonly done by combining a cross-entropy loss with contrastive learning in the form of triplet loss:

$$\mathcal{L}(z, \tilde{y}, y) = \mathcal{L}_{\text{tri}}(z) + \lambda \mathcal{L}_{\text{CE}}(\tilde{y}, y) \tag{1}$$

where $\mathcal{L}_{\text{tri}}$ is a soft-margin triplet loss that uses the object embedding z and $\mathcal{L}_{\text{CE}}$ denotes a cross-entropy loss that uses the output of a linear classifier, $\tilde{y}$. Here, λ acts as a balancing coefficient.

To establish a strong baseline for future multimodal ReID research and investigate the impacts of fusion, we study three separate vanilla late-fusion strategies: fusion by concatenation (*Fusion-concat*), fusion by averaging (*Fusion-avg*), and post-training fusion by unimodal concatenation (*UniCat*).

More specifically, for each modality we train an unshared backbone, f_i that produces a modality-specific embedding, z_i. These modality-specific embeddings can then be combined into a multimodal representation: $z_{\text{fuse}} = \Theta(z_1, z_2, ...z_M)$ where Θ is the chosen fusion operator. The difference among the training strategies lies in whether global or local loss functions are used. In *Fusion-avg* or *Fusion-concat*, we use a global loss function that operates on the fused multimodal representation, z_{fuse}: $\mathcal{L}_{\text{fusion}} = \mathcal{L}(z_{\text{fuse}}, \tilde{y}_{\text{fuse}}, y)$. In this case, the back-propagated gradients for each modality are entangled as described in [9]. Conversely, in *UniCat* we train each modality encoder independently with their own local loss function: $\mathcal{L}_{\text{post}} = \sum_i^M \mathcal{L}(z_i, \tilde{y}_i, y)$ where i is summed over all modalities.

Here, note that each modality's loss is fully disentangled from the others. In this case, fusion occurs only at inference when we concatenate the modality-specific embeddings into the final multimodal representation, z_{fuse}.

3 Results

We benchmark the described late-fusion strategies on three multimodal ReID datasets. RGBNT100 [5] is a multimodal vehicle dataset consisting of coincidental visible (RGB), near-infrared (NIR), and thermal-infrared (TIR) images for 100 vehicles from multiple views as shown in Figure 1. RGBN300 is an extension of RGBNT100 in which (RGB, NIR) pairs are provided for 300 vehicles. Lastly, RGBNT201 [14] is a multimodal person dataset that provides (RGB, NIR, TIR) image triples for 201 different individuals. The key benchmarks, mAP and Rank-1, are informed by the cumulative match curve (CMC) [17]. Notably, mAP offers an overarching performance measure.

Figure 1: RGB, NIR, TIR images from RGBNT100 [5].

We compare our late-fusion baseline strategies with the previous state-of-the-art multimodal ReID models. In particular, we test two of the most widely-used backbones found in other ReID models, ResNet-50 and ViT-B.

In Table 1 we observe that RGBNT100/RGBN300 with *UniCat* and RGBNT201 with *Fusion-concat* do strikingly better than the other best-known fusion approaches, achieving state-of-the-art performance. We observe backbone choice to have a more dramatic impact on RGBNT201 than RBGNT100/RGBN300. Since RGBNT201 contains 2x fewer datapoints than RGBNT100, data-hungry ViTs may struggle to learn with lack of inductive biases compared to CNNs [18]. Overall, we clearly observe fusion strategy as a leading factor on representation quality.

Model	Backbone	RGBNT100		RGBN300		RGBNT201	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
PFNET [14]	ResNet-50	68.1†	94.1†	-	-	31.8	54.6
IEEE [12]	ResNet-50	61.3†	87.8†	-	-	46.42	47.13
HAMNet [5]	ResNet-50	64.1	84.7	61.9	84.0	27.68*	26.32*
CCNet [15]	ResNet-50	77.2	96.3	-	-	-	-
PHT [8]	ViT-B	79.9	92.7	79.3	**93.7**	-	-
Fusion-avg	ResNet-50	75.3 ±0.6	92.2 ±0.9	74.9 ±0.5	91.2 ±0.6	57.6 ±2.9	61.0 ±3.0
Fusion-concat	ResNet-50	75.1 ±0.9	92.7 ±0.6	74.7 ±0.4	91.3 ±0.2	**63.0 ±1.9**	**66.7 ±3.1**
UniCat	ResNet-50	79.4 ±0.5	94.9 ±0.8	76.7 ±0.4	92.3 ±0.4	38.1 ±1.3	33.6 ±1.7
Fusion-avg	ViT-B	76.1 ±0.3	90.1 ±0.9	78.1 ±0.1	91.4 ±0.4	56.1 ±2.2	54.9 ±1.8
Fusion-concat	ViT-B	75.9 ±0.7	92.3 ±0.4	77.4 ±0.3	91.8 ±0.3	61.7 ±2.1	60.6 ±1.2
UniCat	ViT-B	**81.3 ±0.9**	**97.5 ±0.3**	**80.2 ±0.2**	92.9 ±0.4	57.6 ±2.0	58.0 ±1.8

Table 1: Comparison between current state-of-the-art multimodal ReID models with our strong multimodal baselines. Metrics for our models were calcluated by averaging performance across four random seeds. †, $*$ indicates the results were taken from [15] and [12], respectively.

To gain intuition for the dynamics at play, we further compare all of our modality backbones on the task of unimodal ReID. That is, we use our trained modality encoders from Table 1 and evaluate unimodal ReID performance on the test sets. Because the multimodal embeddings in all fusion cases are simply poolings of unimodal embeddings, multimodal ReID performance here is closely tied to the corresponding unimodal ReID performance of the constituent modalities. As shown in Table 2, for RGBNT100 we observe that all unimodal ReID performance has been negatively affected by training with fusion. Surprisingly, we also observe this to be the case in RGBNT201, with the exception of one modality - NIR (see Table 3). We can consider this modality a weak-link since when trained in isolation (*UniCat*) its performance is notably lower than RGB/TIR. As multimodal ReID is most affected by the weakest-link, we deduce that the performance gains of NIR in the case of *Fusion-concat* outweighed the performance drops in RGB/TIR.

We suspect these effects are largely due to the impact and nuances of modality laziness. The results shown by RGBNT100/RGBN300 demonstrate the negative effects that modality laziness can have -

Model	Backbone	RGB		NIR		TIR	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Fusion-avg	ResNet-50	48.6 ±1.4	71.9 ±0.9	40.4 ±0.5	62.3 ±1.3	41.1 ±1.1	71.0 ±2.3
Fusion-concat	ResNet-50	44.7 ±1.2	63.9 ±2.8	27.3 ±1.0	43.5 ±1.1	36.9 ±1.5	62.9 ±1.3
UniCat	ResNet-50	55.0 ±0.8	77.2 ±0.9	46.0 ±1.1	67.3 ±1.0	**52.5 ±0.4**	**82.5 ±1.3**
Fusion-avg	ViT-B	51.7 ±0.8	73.5 ±0.9	41.8 ±0.6	59.4 ±2.7	31.0 ±0.9	53.6 ±2.8
Fusion-concat	ViT-B	50.0 ±4.1	70.8 ±5.6	38.9 ±2.5	60.7 ±1.5	34.1 ±1.6	61.6 ±3.0
UniCat	ViT-B	**58.4 ±0.83**	**80.2 ±1.2**	**51.4 ±0.8**	**71.5 ±1.5**	46.1 ±2.5	76.0 ±3.2

Table 2: RGBNT100 [5] Evaluation of unimodal ReID using the embeddings learned from our multimodal late-fusion ReID baselines.

Model	Backbone	RGB		NIR		TIR	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Fusion-avg	ResNet-50	21.4 ±1.8	17.5 ±3.0	14.8 ±1.0	11.5 ±0.6	33.0 ±1.9	33.6 ±3.5
Fusion-concat	ResNet-50	16.7 ±2.3	14.2 ±2.8	18.9 ±2.1	16.6 ±2.8	28.5 ±0.6	30.0 ±1.5
UniCat	ResNet-50	23.2 ±2.0	18.34 ±2.5	11.5 ±0.6	7.4 ±0.3	36.0 ±0.3	36.9 ±1.2
Fusion-avg	ViT-B	28.1 ±1.8	25.2 ±1.6	20.5 ±2.2	16.5 ±3.0	27.6 ±0.8	26.2 ±1.1
Fusion-concat	ViT-B	28.2 ±2.7	25.8 ±2.8	**24.7 ±1.5**	**23.7 ±0.9**	30.0 ±0.8	29.0 ±1.5
UniCat	ViT-B	**31.9 ±0.8**	**29.6 ±1.2**	20.9 ±1.1	16.8 ±1.5	**37.5 ±1.3**	**36.2 ±1.6**

Table 3: RGBNT201 [14] Evaluation of unimodal ReID using the embeddings learned from our multimodal late-fusion ReID baselines.

leading to underutilized unimodal features. Conversely, for datasets such as RGBNT201 where noisy, less-reliable modalities exist, modality laziness may offer a training relief by helping to prevent those modalities from resorting to high-degrees of overfitting. This latter notion has also been hypothesized by Liu et al. [6].

3.1 RGBNT100 Training Performance

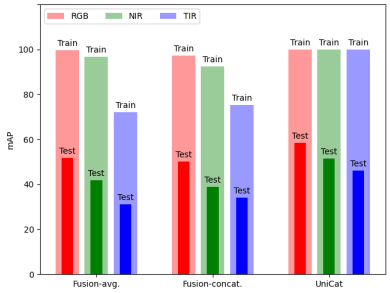

Figure 2: Comparison of train and test unimodal ReID performance for our late-fusion mulitmodal baselines (ViT-B) on RGBNT100.

To further distinguish whether modality laziness or general overfitting is at play in the performance drop observed when training RGBNT100 with fusion, we evaluate the unimodal ReID performance (given by mAP) of the train set for RGBNT100 (see Figure 2). Mirroring the test behavior, we see that the unimodal mAP for the train set was inhibited for *Fusion-avg* and *Fusion-concat* compared to *UniCat*. Thus, contrary to the hypothesis of [11], the performance decline that we observed when training with fusion (Table 1) can be primarily linked to modality laziness, and it is actually *UniCat* that appears to be more at risk for overfitting to the train data. This also helps to demonstrate how the impacts of modality laziness may at times be beneficial for weaker, less-informative modalities that are especially overfitting-prone, as we observed in RGBNT201 (Tables 1 and 3).

3.2 Effects of Fusion Strategies in Ensemble Learning

Peng et al. [9] suggests that modality laziness occurs when some modalities are more dominant than others. However, we suspect that this phenomenon can occur even when modalities have comparable,

Model	RGBNT100						Market1501	
	RGB		NIR		TIR			
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Fusion avg	57.7	79.1	48.6	67.7	48.6	76.9	87.6	94.9
Fusion-concat	58.4	81.0	48.9	70.0	47.4	80.8	87.9	94.8
UniCat	**60.4**	**84.4**	**50.7**	**72.9**	**49.7**	**81.1**	**89.1**	**95.4**

Table 4: Comparison of late-fusion implementations for a two-backbone (ViT-B) unimodal ensemble learner.

or even equal, predictive strengths. To demonstrate this idea, we look to ensembles. Specifically, we train an ensemble of unimodal ReID models on a single modality. When training ensembles, it is common practice to train each ensemble member individually and fuse their predictions only at inference (i.e. bagging). We compare this standard strategy of training each ensemble member independently (*UniCat*) with training them jointly by having a global loss use their fused output embedding (*Fusion-avg/concat*).

For this experiment, we investigate each modality within RGBNT100 as well as the popular unimodal person ReID dataset, Market1501 [16]. We choose ViT-B as our model backbone and use ImageNet pre-trained weights. The classifiers and bottlenecks are randomly instantiated via [4], so that we may fully expect the backbones to have unique final solutions.

As shown in Table 4, there is a boost in unimodal RGBNT100 ReID performance for all ensemble models when compared to Table 2, which is expected as ensembling helps to diminish the impacts of overfitting. However, what is notable is that in all unimodal tasks, training the ensemble using a global loss (*Fusion-avg/concat*) resulted in a lower performance than would otherwise have been obtained if each member were trained independently (*UniCat*). This underscores the fundamental nature by which modality laziness operates and how the phenomenon should be considered even when training with balanced modalities.

4 Discussion

Our results indicate that when different modalities are jointly optimized (fusion training), they may not capture all available task-relevant information, leading to potential underrepresentation. This is especially pronounced in datasets like RGBNT100/RGBN300. But there's a silver lining: while certain modalities suffer, noisy ones may benefit since joint training may act as a form of regularization for such modalities, preventing overfitting to task-irrelevant data. Thus, a delicate equilibrium emerges in multimodal ReID learning. We find that the *UniCat* approach using a ViT-B backbone outperforms other models in both RGBNT100 and RGBN300. Here, each modality—RGB, NIR, and TIR—benefits from the pressure of being trained in isolation. Challenging views like those in [11] that focus on over-parameterization, our results suggest that joint optimization's nuances may limit modalities from capturing complete task-relevant information, therein lying the negative effects of modality laziness. Conversely, in certain scenarios, such as with the RGBNT201 dataset, the utility of modality laziness emerges. Specifically, the *Fusion-concat* method bolsters the performance of NIR—a notably weaker modality. While NIR struggles when optimized individually, joint optimization uplifts its performance. We attribute this to modality laziness acting as a shield against the modality's inherent noise, resonating with the observations made by [6]. This not only highlights the protective role of modality laziness but also underscores its potential as a regularization mechanism, especially for modalities with less task-relevant information. Previous solutions to modality laziness ([2], [9]) do not consider the full scope of the phenomenon.

Our findings highlight the need for a unified fusion approach to address the multifaceted nature of modality laziness and show that much research is still needed to inform more holistic, nuanced multimodal frameworks. We encourage researchers in this field to validate simple post-training fusion baselines like *UniCat* first in their setting, before pursuing more complex fusion strategies. Lastly, frameworks that are both model parameter efficient (e.g., a shared backbone encoder) and data efficient are essential for future architecture scalability.

References

[1] Cem Akkus, Luyang Chu, Vladana Djakovic, Steffen Jauch-Walser, Philipp Koch, Giacomo Loss, Christopher Marquardt, Marco Moldovan, Nadja Sauter, Maximilian Schneider, Rickmer Schulte, Karol Urbanczyk, Jann Goschenhofer, Christian Heumann, Rasmus Hvingelby, Daniel Schalk, and Matthias Aßenmacher. Multimodal deep learning, 2023.

[2] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning, 2023.

[3] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all, 2023.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

[5] Hongchao Li, Chenglong Li, Xianpeng Zhu, Aihua Zheng, and Bin Luo. Multi-Spectral Vehicle Re-Identification: A Challenge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11345–11353, April 2020.

[6] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning, 2018.

[7] Purvanshi Mehta. Multimodal deep learning, 2020.

[8] Wenjie Pan, Linhan Huang, Jianbao Liang, Lan Hong, and Jianqing Zhu. Progressively hybrid transformer for multi-modal vehicle re-identification. *Sensors*, 23(9), 2023.

[9] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation, 2022.

[10] Dhanesh Ramachandram and Graham W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[11] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard?, 2020.

[12] Zi Wang, Chenglong Li, Aihua Zheng, Ran He, and Jin Tang. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2633–2641, Jun. 2022.

[13] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *CoRR*, abs/2001.04193, 2020.

[14] Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:3529–3537, 05 2021.

[15] Aihua Zheng, Xianpeng Zhu, Zhiqi Ma, Chenglong Li, Jin Tang, and Jixin Ma. Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark. *Information Fusion*, 100:101901, 2023.

[16] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, Los Alamitos, CA, USA, dec 2015. IEEE Computer Society.

[17] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[18] Haoran Zhu, Boyuan Chen, and Carter Yang. Understanding why vit trains badly on small datasets: An intuitive perspective, 2023.