GraFT: Gradual Fusion Transformer for Multimodal Re-Identification

Anonymous WACV Algorithms Track submission

Paper ID 2577

Abstract

Object Re-Identification (ReID) is pivotal in computer vision, witnessing an escalating demand for adept multimodal representation learning. Current models, although promising, reveal scalability limitations with increasing modalities as they rely heavily on late fusion, which postpones the integration of specific modality insights. Addressing this, we introduce the ***Gradual Fusion Transformer (GraFT)*** *for multimodal ReID. At its core, GraFT employs learnable fusion tokens that guide self-attention across encoders, adeptly capturing both modality-specific and object-specific features. Further bolstering its efficacy, we introduce a novel training paradigm combined with an augmented triplet loss, optimizing the ReID feature embedding space. We demonstrate these enhancements through extensive ablation studies and show that GraFT consistently surpasses established multimodal ReID benchmarks. Additionally, aiming for deployment versatility, we've integrated neural network pruning into GraFT, offering a balance between model size and performance. Most recent state-of-the-art multimodal ReID methods are not reproducible nor readily validated. To address this gap, we release our codebase to showcase a new state-of-the-art in reproducible multimodal ReID:* https://anonymous.4open.science/r/GraFT/

1. Introduction

Object re-identification (ReID) is the computer vision task of determining whether an object of interest has appeared previously at a distinct place and/or time. At its core, ReID is a matching task, wherein a sampled *query* image is contrasted against a pre-existing *gallery* of images. This task has significant applications in areas such as retail, robotics, multimedia, and surveillance. However, ReID comes with significant challenges since the captured representations of objects are subject to a range of uncertainties such as different sensor viewpoints, object poses, occlusions, varying low-resolutions, and environmental conditions [40]. Furthermore, because most conventional ReID algorithms are designed to operate on conventional visible spectrum Red, Green, Blue (RGB) images, there are formidable challenges in less-than-ideal environmental scenarios such as low-light or hazy conditions, similar to the limitations of the human eye. To address these issues, additional sensor modalities such as those on the infrared spectrum are commonly used to complement the RGB sensors. However, in the context of multimodal ReID, the critical problem becomes the effective fusion of such diverse data representations, where learning useful object features and the nuances of each modality is essential.

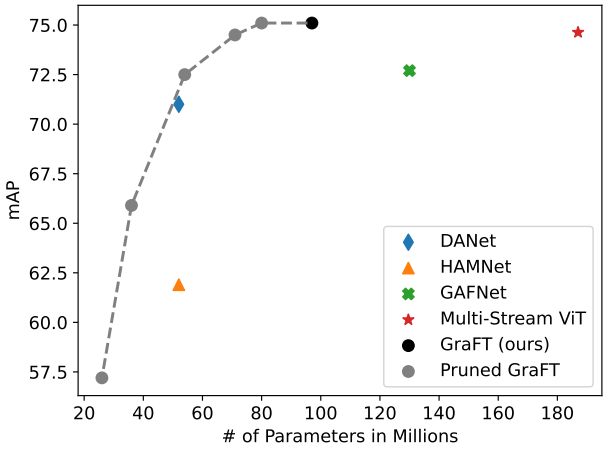

Figure 1. Model Size vs Performance on RGBN300 Benchmark

Deep learning model architectures for multimodal ReID generally fall into two categories: early fusion and late fusion. Early fusion involves the concatenation of images from different modalities, which are then jointly processed through the model. This aims to allow for a more combined understanding of the scene but has the trade-off of sacrificing modality-specific information. Consequently, the richness and depth that each modality offers might not be harnessed to its fullest potential. In contrast, late fusion involves processing each modality individually, and subsequently combining the respective embeddings towards the

output of the model. This approach ensures the modality-specific information is independently learned, but may sacrifice object-specific understanding and suffer from impractical model size increases a new modalities are added.

To address the limitations inherent to both early and late fusion, we propose an effective multimodal learning solution that takes a *gradual* approach to fusion to preserve both modality and object-specific features throughout the model. Our core contributions can be summarized as follows:

- We propose a **Gradual Fusion Transformer (GraFT)** architecture for multimodal ReID that uses learnable fusion tokens to guide self-attention across encoder layers to extract modality *and* object-specific features.
- We develop a unique combination of training paradigms including an augmentation to triplet loss for a more effective ReID feature embedding space.
- We conduct extensive experiments and ablation studies on GraFT using the multimodal ReID benchmark datasets RGBNT100 and RGBN300 [20] to outperform existing methods as seen in Fig. 1.
- To maximize deployment flexibility, we integrate a neural network pruning capability to allow for a variety of GraFT model size and performance options.
- Most recent state-of-the-art multimodal ReID methods are not reproducible nor readily validated. To address this gap, we release our codebase to showcase a new state-of-the-art in reproducible multimodal ReID.

2. Related Works

2.1. Re-Identification

In recent years, deep learning has rapidly pushed unimodal RGB ReID to new state-of-the-art levels [5, 9, 15, 22, 31, 32, 36, 50] achieving impressive matching accuracy in constrained settings. However, despite this success, relying solely on RGB imagery has inherent weaknesses that present opportunities to explore multimodal techniques. RGB lacks invariance to variations in lighting, occlusion, and viewpoint that commonly occur in uncontrolled real-world ReID scenarios [19, 29, 47]. Furthermore, visibility is significantly degraded under nighttime conditions where illumination is limited [43]. In contrast, near-infrared (NIR) and thermal-infrared (TIR) imaging can provide invariant geometric identity cues highly valuable in low light settings [12, 20, 33] as shown in Fig. 2. Although a handful of studies have combined RGB and infrared by first individually processing each modality and then concatenating the results together [3, 17, 20, 41], these approaches do not deeply integrate the complementary modalities architecturally. Thus, while unimodal RGB ReID is mature, ample untapped opportunities remain to overcome unimodal limitations by developing principled multimodal fusion approaches. Our proposed GraFT method aims to address these gaps by learning optimized fusion and modality-specific representations to integrate multiple complementary cues for enriched ReID.

(a) RGB (b) NIR (c) TIR

Figure 2. Example images from the RGBNT100 dataset [20].

2.2. Multimodal Representation Learning

Multimodal representation learning aims to integrate information from different data modalities (e.g., images and text) into a joint embedding space. Convolutional neural networks (CNNs) and Vision Transformers (ViTs) are commonly used as visual encoders for this task [25]. CNNs apply convolutional filters to hierarchically extract visual features from images while ViTs split images into patches and leverage self-attention, allowing modeling of long-range dependencies.

The key challenge is determining how to effectively combine the unimodal representations from each modality into the joint space. Early fusion approaches directly concatenate the raw inputs from each modality before passing them to a joint model. This enables learning cross-modal interactions and aims to create integrated multimodal representations. However, it lacks flexibility since all modalities are handled identically. Late fusion first encodes each modality separately with customized encoders, before concatenating their outputs. This allows architectural optimization tailored to each modality. However, it can miss important joint representations by delaying fusion [2, 42]. Recent methods aim to get the benefits of both approaches through techniques like specialized attention mechanisms or gating to dynamically modulate fusion based on the specific context [13, 26, 30]. These specialized fusion methods are often designed for particular downstream tasks like visual question answering [6] or classification [13]. However, they do not focus on learning a general joint embedding space that works well for tasks like multimodal person or vehicle re-identification, which is what our proposed solution provides. Our approach focuses on getting a high-quality joint representation for ReID by fusing modalities in a principled manner.

2.3. Multimodal Re-Identification

Although the additional of modalities can improve the robustness and performance of ReID, a core challenge remains: how to effectively merge these different data types. Current research suggests that many existing techniques either combine the data too early or too late, compromising the results [14, 17, 18, 20, 27, 37, 44, 45]. While much research has been dedicated to merging standard RGB visual data [15, 22, 39], there's a clear need for methods that can seamlessly integrate additional visual modalities such as depth and infrared for enhanced identification.

Our approach aims to fill this gap by learning discriminative yet robust embeddings optimized for jointly representing multimodal cues for ReID. We build on top of prior fusion insights and propose innovations to create purpose-built embeddings that efficiently fuse modality-specific information into object-specific features. In particular, we utilize gradual fusion to carefully control the flow of information through transformers and modality encoders to produce robust object representations while maintaining useful specific information from each modality. By tackling representation learning for efficient and holistic fusion, our method represents a significant advance. Our experiments demonstrate state-of-the-art performance on benchmark multimodal ReID datasets, highlighting the benefits of joint embeddings tailored for unified multimodal matching. The powerful yet efficient embeddings produced by our approach could enable deployment of multimodal reID systems in real-world applications.

3. Method / GraFT Fusion Technique

In this section, we describe our proposed method depicted in Fig. 3: GraFT. We first give a high-level overview of our method, then briefly define the popular Vision Transformer (ViT), which serves as a backbone feature extractor. Then, we discuss our token fusion technique, motivate its usage for constructing efficient embedding spaces for Vehicle ReID, and explain in detail the flow of the network from input to output.

3.1. Method Overview

To accomplish gradual fusion, we carefully restrict the flow of information through our model to simultaneously (a) produce a robust object representation and (b) maintain useful and specific information from each modality. First, a transformer backbone extracts features from our raw data, creating information-rich data tokens. We then introduce a learnable fusion token, which is joined with the data tokens and passed separately through each modality's corresponding modality encoder. Through cross-attention, each modality encoder embeds information unique to its respective modality within the fusion token. Finally, the fusion token recombines into one robust object representation through averaging.

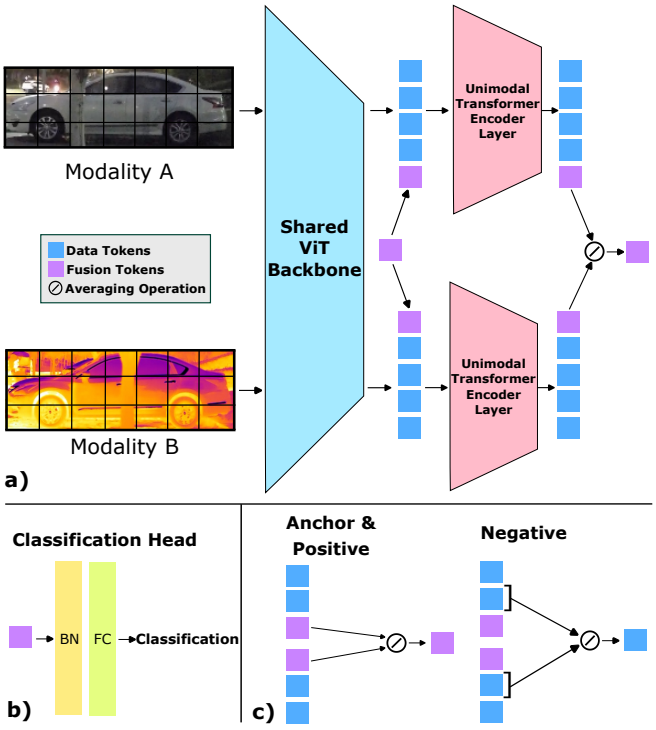

Figure 3. GraFT model architecture. a) Main architecture depicting two modality inputs without loss of generality to more modalities. GraFT leverages learnable fusion tokens to gradually fuse together multimodal information. b) The final fusion token embedding is passed to the BNNeck [24] and a classifier Fully Connected (FC) layer. c) Our model is trained via a contrastive loss where the fusion tokens are used as the anchor and positive samples and a random data token is used as the negative sample.

3.2. Backbone

The Vision Transformer (ViT) [7] is a derivative of the Transformer architecture [35]. It adapts the original Transformer's architecture for images by treating them as a sequence of fixed-size patches, equivalent to the tokens in text data. These patches then undergo the same self-attention and feed-forward network operations. In our study, we first use a pretrained ViT-B model on the ImageNet dataset as the feature extractor for ReID tasks, in line with other transformer-based ReID works [15]. Empirically, we find that DeiT-B model [34] works the best due to its data distillation pretraining scheme due to a similar data limitation problem, so we adopt that as our final backbone.

Next, the features extracted from the Vision Transformer (ViT) backbone for each modality are processed. These data tokens are then fed into their corresponding unimodal transformer encoder layers, in conjunction with learnable fusion tokens that are explained in the following section. The final

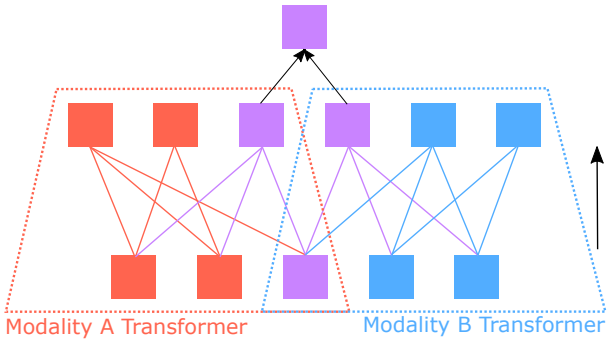

Figure 4. Unimodal transformer encoder layer attention flow visualization. Fusion tokens act as the bridge between modalities with modality information being gradually fused through attention.

fusion token is then routed to the classification head.

3.3. Learnable Fusion Tokens

To encourage concise communication between features from multiple modalities during fusion while leveraging the self-attention capabilities of Transformers, we employ a learnable fusion token, $\mathbf{T_f}$ created with Xavier initialization [11]. We restrict the flow of attention between modalities solely through the fusion token as seen in Fig. 4, compelling the token to learn modality-agnostic features at the intersection of all modalities. This method reduces the computational overhead of attention since attention is only required between the fusion tokens and the input sequences from each modality rather than across all modalities.

Additionally, by averaging the transformed outputs for each modality encoder's fusion token, the amount of fusion token parameters required remains constant regardless of the number of modalities, thereby ensuring scalability without compromising fusion proficiency. We also note that similar methods of fusion through learnable tokens have achieved state-of-the-art results on audio-visual discriminative tasks [26], but this idea is novel to the ReID task, one that requires the generation of well-ordered embeddings in the vector space.

3.4. Architecture Walkthrough

Given M-many modalities, our inputs are tuples of coincidental images for each modality: $(x_1, x_2, ...x_M)$ such that for each image, $x_i \in \mathbb{R}^{C\times H\times W}$ for $i \in [M]$. Each image is separately fed through a patch embedding that splits the image into $L_d = HW/P^2$ many non-overlapping patches of shape P^2C, such that (P, P) is the resolution of each patch. We use $P = 16$ for our base model. To create the patch embeddings, each patch is linearly projected from shape P^2C to some latent vector size D. The shared backbone takes the patched images from each modality and generates useful features as the data tokens, $\mathbf{T_{d_i}} \in \mathbb{R}^{L_d\times D}$ for some modality i.

$$\text{Patch Embed} : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{L_d\times D} \quad (1)$$

$$\text{Backbone} : \mathbb{R}^{L_d\times D} \rightarrow \mathbb{R}^{L_d\times D} \quad (2)$$

The data and learnable fusion tokens are then concatenated along the sequence length dimension and passed into the modality encoders. The encoders each consist of a multiheaded self-attention module (MHA), and a multilayer perceptron (MLP). For modality $i \in [M]$, we have

$$\mathbf{E_i^1} = \mathit{Concat}(\mathbf{T_f^*}, \mathbf{T_{d_i}}) \quad (3)$$

$$\mathbf{E_i^2} = \mathit{MLP}(\mathit{MHA}(\mathbf{E_i^1}; \theta_\mathbf{i})) \quad (4)$$

$$\mathbf{Z_{f_i}}, \mathbf{Z_{d_i}} = \mathit{Split}(\mathbf{E_i^2}) \quad (5)$$

such that

$$\mathbf{E_i^1}, \mathbf{E_i^2} \in \mathbb{R}^{(L_d+1)\times D}, \mathbf{Z_{f_i}} \in \mathbb{R}^{L_f\times D}, \mathbf{Z_{d_i}} \in \mathbb{R}^{L_d\times D}$$

Note that $\mathbf{T_f^*}$ signifies the expansion or sharing of the fusion token weights at every modality. We then get an aggregate of the fusion tokens from each modality through averaging,

$$\mathbf{Z_f} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{Z_{f_i}}. \quad (6)$$

The final embedding for each sample is just the fused token,

$$\boldsymbol{Embed} = \mathbf{Z_f} \quad (7)$$

are subsequently forwarded to the ReID task to calculate distance metrics. During training, ***Embed*** is also passed to a batch normalization (BN) and fully connected linear layer (FC), both with bias turned off to perform ID classification [24]. Applying BN to the embeddings transforms the feature space into a hypersphere centered at the origin with mean zero and unit variance. This standardized distribution aligns with the assumptions of linear models, enabling more effective separation. The resulting spherical input simplifies downstream classification by removing covariate shifts and scaling imbalances between dimensions. However, as only the bias terms of BN and the FC are frozen in our proposed approach, the weights remain trainable via backpropagation. Thus, the model retains some adaptability while benefiting from the normalized feature distribution.

4. Training Paradigms

4.1. Frozen to unfrozen backbone

To optimally leverage a backbone pretrained on RGB datasets like ImageNet, we initially employ a frozen pretrained DeIT-B as a general feature extractor. This step encourages the modality-specific encoders and fusion tokens to learn modality-specific features upon general ones and

also compensates for the limited training samples available in most ReID datasets. Once the modality encoders are adequately trained, we proceed to fine-tune the general feature extractor to more accurately represent each modality's distribution. We divide training into two stages because using a higher learning rate in the initial stage ensures that the untrained parameters can effectively explore the search space.

4.2. Contrastive Loss

Strong clustering within the feature embedding space is an essential characteristic of successful ReID models. To encourage clustering, we employ contrastive loss both between triples and over the entire embedding space.

Between individual triples, which consist of an "anchor" identity, a "positive" match, and a "negative" non-match, we employ soft margin triplet loss [16]. Formally, given the triple $\{anchor, positive, negative\}$ with feature embeddings $\{\mathbf{f_a}, \mathbf{f_p}, \mathbf{f_n}\}$, soft margin loss $\mathcal{L}_T$ is as follows:

$$\mathcal{L}_T = \log(1 + \exp(\|\mathbf{f_a} - \mathbf{f_p}\|_2^2) - \|\mathbf{f_a} - \mathbf{f_n}\|_2^2)) \quad (8)$$

Over the whole set of feature embeddings, we employ center loss to penalize each embedding's distance from learnable ID-based centroids [38]. Given the feature embeddings for our anchor identities, $\mathbf{f_a}$, we compute the cosine distance between each embedding and the learned centroid $\mathbf{c_y}$ of its corresponding ID y. This computation is performed over the minibatch as follows to compute the center loss $\mathcal{L}_C$, where B is batch size. We formalize this as

$$\mathcal{L}_C = \sum_{j=1}^{B} \|\mathbf{f_{a_j}} - \mathbf{c_{y_j}}\| \quad (9)$$

Our complete loss function for ReID combines triplet loss, center loss, and cross-entropy loss with vehicle ID's as classes, $\mathcal{L}_{CE}$, via weighted sum:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_T + \beta\mathcal{L}_C + \gamma\mathcal{L}_{CE} \quad (10)$$

For training with a frozen versus a frozen feature extractor, we found best results with $\alpha = 0.5, \beta = 0, \gamma = 0.5$ and $\alpha = 0.5, \beta = 0.0005, \gamma = 0.5$, respectively.

4.3. Augmented Triplet Loss

For the anchor and positive ID, we simply use the fused tokens $\mathbf{Z_{f_a}}$ and $\mathbf{Z_{f_p}}$ as the feature embeddings inputted to triplet loss ($\mathbf{f_a}$ and $\mathbf{f_p}$). However, since the fusion tokens are a subset of the data tokens semantically, to encourage effective clustering within classes and optimize for a more linearly separable latent space, we utilize a subset of the data tokens $\mathbf{Z_{d_i}}$ from each modality encoder output $\mathbf{E}_i^2$ as the negative embedding $\mathbf{f_n}$. For each modality, we take a sample token from an arbitrary fixed index (e.g., index 0):

$$\mathbf{Z}_{d_{i_0}} = Sample(\mathbf{Z}_{d_i}) \quad (11)$$

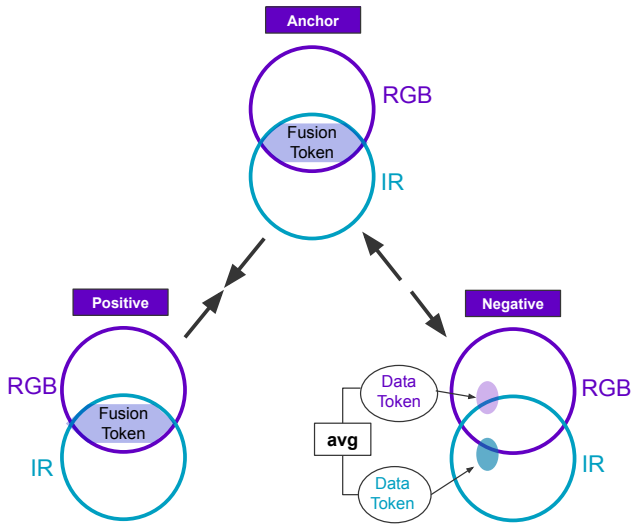

Figure 5. Augmented triplet visual explanation using sets. Without loss of generality, RGB and IR represent modality A and B. Fusion tokens act as the intersection of modalities while data tokens are the union of modalities. Through augmented triplet training, anchor and positive samples have their modality intersections pulled together while anchor and negative have the modality intersection pushed away from the modality union.

As with the fusion tokens, we then average the data samples across modalities to construct our negative input for triplet soft margin loss, $\mathbf{f_n}$:

$$\mathbf{f_n} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{Z_{d_{i_0}}} \quad (12)$$

We denote augmented triplet loss as Fusion-Fusion-Data (*FFD*), indicating which token is used as the input to anchor positive, and negative, respectively. We similarly define other possible combinations; for example, standard triplet loss, with all samples having fusion token inputs, would be Fusion-Fusion-Fusion (*FFF*).

The core benefit of augmented triplet loss lies in the objective of triplet loss: minimizing the distance between the anchor and the positive embedding while maximizing the distance between the anchor and the negative embedding. Directly using the fusion token for the anchor and positive samples in triplet loss awards the design of similar feature embeddings for two samples of the same ID. Meanwhile, using data tokens for the negative ID captures a broader distribution of the negative samples. Essentially, as seen in Fig. 5, augmented triplet loss leverages the modality intersection to optimize for similar anchor and positive feature embeddings, while leveraging the modality union to further differentiate anchor and negative feature embeddings.

We show empirically in Section 6. that the specific index

sampled has no significant impact on model performance, and that reintroducing the data token in the negative ID has a significant positive impact on performance.

5. Results

5.1. Datasets Details

We train and evaluate our Vehicle ReID method using the benchmark RGBNT100 and RGBN300 dataset [20]. The RGBNT100 dataset consists of coincidental RGB, NIR, and TIR coincidental images of 100 unique vehicle IDs from various different camera views. Similarly, the RN300 dataset consists of coincidental RGB and NIR images of 300 unique vehicle IDs from different perspectives. For RGBNT100, there are a total of 51,750 images with 8,675 in train, 1,715 in query, and 8,575 in *gallery*. For RGBN300, there are a total of 100,250 images with 25,200 in train, 4,985 in *query*, and 24,925 in *gallery*. We sampled triples from the training data randomly and generate multiple sets of triples for every unique image in the dataset to broaden the scope of contrastive loss. To further increase the number of training instances, we sampled multiple positive examples for every unique image sample and set that as a hyperparameter. We found that doing this drastically improved training speed.

5.2. Implementation Details

For software tooling, we use the common deep learning framework PyTorch 2.0.1 [28], CUDA 11.6, and Python 3.8. For hardware, we use a cluster of eight A6000 GPUs for distributed training leveraging PyTorch Fabric Lightning [8] with the Data Distributed Parallel (DDP) protocol [21]. For the data, we first resize the images to 224x224 pixels then apply horizontal/vertical flips and random erasing [48]. As explained in Section 5.1, we also sample 8 positive examples per anchor to accelerate training. Each mini-batch contains 26 (max amount in GPU virtual memory) anchor, positive, and negative triplets of the respective paired modalities, which are RGB, NIR, and TIR for RGBNT100 and RGB and NIR for RGBN300. For optimization, we utilize the AdamW optimizer [23] with learning rate and weight decay hyperparameter tuned via optuna [1]. For the loss function, we found the best loss hyperparameters to be $\alpha = 0.5, \beta = 0.005$, and $\gamma = 0.5$. For the architecture, We use only one fusion token to bottleneck the information flow such that the classifier head does not overfit and use one Transformer Encoder Layer provided by the PyTorch library for each unimodal transformer encoder layer.

Inspired by [4], we leverage two-stage training. For the **first stage**, we remove center loss as there are no centroids initially to reduce intra-class variation. Through empirical observation and intuition, we find that including the center loss in stage one only slows down training and leads to sub-optimal convergence. During stage one training, we also freeze the shared pretrained ViT backbone to only train the unimodel transformer to only train unimodal transformer encoder layers from scratch since those would have large initial losses and take larger steps than the pretrained model would be suited for. To optimizer stage one, we use a learning rate scheduler with linear warmup, square root decay, and finally linear cooldown for the last 10% of training. For the **second stage**, we unfreeze the shared pretrained ViT backbone to finetune the entire model and add center loss back in to the overall loss function as there exist class clusters optimized by our augmented triplet loss. For training stage two, we set label smoothing to 0.1 similar to [24] and use a constant learning rate at 5e-6.

Methods	Params	mAP (%)	R1 (%)	R5 (%)	R10 (%)
HAMNet [20]	78M	65.4	85.5	87.9	88.8
DANet [18]	78M	N/A	N/A	N/A	N/A
GAFNet [14]	130M	74.4	93.4	94.5	95.0
Multi-Stream ViT	274M	74.6	91.3	92.8	93.5
GraFT (Ours)	101M	**76.6**	**94.3**	**95.3**	**96.0**

Table 1. ReID performance comparison of GraFT and other methods on the RGBNT100 dataset.

Methods	Params	mAP (%)	R1 (%)	R5 (%)	R10 (%)
HAMNet [20]	52M	61.9	84.0	86.0	87.0
DANet [18]	52M	71.0	89.9	90.9	91.5
GAFNet [14]	130M	72.7	91.9	93.6	94.2
Multi-Stream ViT	187M	73.7	91.9	94.1	94.8
GraFT (Ours)	97M	**75.1**	**92.1**	**94.5**	**95.2**

Table 2. ReID performance comparison of GraFT and other methods on the RGBN300 dataset.

5.3. Benchmark Comparisons

We compare the performance between our gradual fusion methods and other reproducible state-of-the-art (SOTA) results in Tables 1 and 2. These comparison works focus exclusively on the task of multimodal vehicle ReID with specialized architectures tackling each part of this difficult task. These benchmarks all optimize the metrics of R1, R5, and R10, which come from isolated points on a cumulative match curve (CMC). As seen in [46], mAP is an aggregate indicator across the CMC curve and indicates whether one algorithm is better at the ReID task as a whole compared to other algorithms.

More specifically, we compare against CNN-based works such as HAMNet [20], DANet [18], and GAFNet

Parameters	mAP↑	R1↑	R5↑	R10↑
101M	76.6	94.3	95.3	96.0
76M	76.4	94.0	94.4	94.6
70M	75.1	94.1	94.9	95.5
60M	73.3	92.2	93.2	94.2
46M	69.0	90.2	91.1	91.7

Table 3. Performance of pruned GraFT variants on RGBNT100. The original dense model has 101 million parameters: one backbone with 3 modality encoders.

Parameters	mAP↑	R1↑	R5↑	R10↑
97M	76.6	94.3	95.3	96.0
80M	75.1	91.0	91.8	92.4
71M	74.7	90.7	91.53	92.1
54M	74.5	90.1	91.1	91.8
45M	72.48	88.67	90.0	91.0
36M	65.9	84.4	84.4	86.1
24M	56.2	77.7	79.2	80.4

Table 4. Performance of pruned GraFT variants on RGBN300. The original dense model has 97 million parameters: one backbone with 2 modality encoders.

[14]. We also compare against a multi-stream ViT baseline, reproduced from [27], where each modality has its own pretrained ViT-B (as opposed to sharing one) and modality-specific embeddings are fused via averaging to attain the final fused token.

5.4. Model Size Analysis

After the creation of the base GraFT model, we study the performance across different model sizes through the use of neural network pruning [10]. The aim of neural network pruning is to remove any superfluous parameters, such as individual weights in our scenario, and maintain performance. Multimodal ReID architectures are large vision models that prove to be difficult for inference on smaller devices. Our goal is to create a flexible model for training, yet compressible for inference. We employ magnitude pruning [49] with finetuning at various sparsities to explore how our model performs under model size constraints. More specifically, we performed few-shot iterative magnitude pruning on the backbone architecture: pruning weights for each query, key, and value along with the projection layers and MLPs.

We prove to be scalable yet compressible, even achieving state-of-the-art performance at a lower parameter count compared to other SOTA as seen in Fig. 1. As shown in Tables 3 & 4, our model collapses after compressing it more than 2.5 times smaller. Through fusion and pruning, we are able to deploy GraFT on more realistic hardware for vehicle ReID in the wild, something not particularly feasible for transformer-based multi-stream approaches [27].

Fusion Method	mAP↑	R1↑	R5↑	R10↑
Vanilla Fusion: CLS Token	60.3	82.1	79.8	83.2
Vanilla Fusion: Averaged Token	62.0	83.9	85.9	87.4
GraFT: Fusion Token (Ours)	**76.4**	**94.0**	**94.4**	**94.6**

Table 5. Performance of current vanilla fusion approaches compared with GraFT's Fusion Technique on RGBNT100

6. Ablations

We conduct a set of ablation studies to assess the influence of our architectural decisions, training strategies, and hyperparameters, as well as to evaluate the scalability of GraFT across different modalities.

In our first ablation study, we compare different fusion techniques on the RGBNT100 dataset as shown in Table 5. In this method, features extracted from each modality are concatenated along the sequence length dimension, complemented by a CLS token. These are subsequently introduced to the Transformer encoder layer. For the subsequent decoder, the aggregate CLS token is utilized. In addition, we evaluate a fusion variant that computes the average of all aggregated feature tokens, analogous to the averaging operation we employ for fusion tokens in Equation 6. All three fusion techniques underwent training under identical conditions and stages. As shown in Table 5, our GraFT fusion approach achieves a notable improvement of 16.1 mAP over the conventional fusion and bests the averaging token fusion method by 14.4 mAP. It is noteworthy that the vanilla fusion averaged token method outperforms the vanilla fusion CLS token technique, indicating that an equally weighted linear combination of the aggregated input yields more useful representations than leveraging a CLS token as an aggregate. We postulate that the sub-optimal performance of vanilla fusion stems from its attempt to fuse all features of all modalities simultaneously, lacking a mechanism to enforce inter-modality coordination. This potentially fails to distill modality-agnostic information crucial for the ReID task into a succinct representation. In contrast, the **GraFT learnable fusion token provides a dynamic mechanism that allows for adaptive inter-modality interactions, which leads to a more efficient and effective representation of the crucial object-specific information for ReID**.

Our second ablation study investigates how the GraFT framework scales with the addition of modalities. We first start by looking at each modality separately in a unimodal GraFT setting and then continue by testing different combinations of modalities beyond just the traditional RGB set-

Modality	mAP↑	R1↑	R5↑	R10↑
T	38.6	63.5	71.0	74.7
N	39.6	60.4	64.2	66.5
R	49.3	70.3	75.0	77.4
R+N	54.3	76.4	79.6	81.6
N+T	62.7	86.8	89.1	90.3
R+T	67.0	89.2	91.3	92.7
R+N+T	**76.6**	**94.3**	**95.3**	**96.0**

Table 6. Performance of our method across different modalities on the RGBNT100 dataset. T=Thermal-IR, N=Near-IR, R=RBG.

Modality	mAP↑	R1↑	R5↑	R10↑
N	51.3	74.1	77.2	79.5
R	60.4	81.5	83.9	86.8
R+N	**75.1**	**91.4**	**92.3**	**92.9**

Table 7. Performance of our method across different modalities on the RN300 dataset. T=Thermal-IR, N=Near-IR, R=RBG.

ting. As anticipated, as more modalities are added, the performance of GraFT improves as shown in Tables 6 and 7 across all performance metrics and the benchmark datasets. This highlights that the **GraFT approach is able to learn representations across modalities in such a way that allows for object and modality specific data characteristics to be efficiently leveraged**.

In our third ablation study, we compare various fusion and data token combinations in order of anchor, positive, and negative as related to Fig. 5. Out of all the combinations, we find that the *FFD* contrastive loss scheme most effectively improves performance by a large margin as shown in Table 8, which follows the intuition as explained in Section 4.3. This demonstrates that our **augmented triplet loss leads to higher performance than triplet loss with standard inputs**, or with any other combination of data token and fusion token inputs. Taking a closer look, only sampling data tokens (*DDD*) outperforms only using fusion tokens (*FFF*), which can be explained since the *DDD* has a much larger distribution of values to conduct contrastive learning on as compared to the averaged *FFF* fusion tokens. However, it is important to note how incorporating fusion tokens in some form (see *FDD, DFF, DFD*) all outperform the *DDD* approach, which indicates that fusion tokens play an important role in our usage of augmented triplet loss for contrastive learning in multimodal ReID.

As a final observation, we experimented with sampling the input for augmented triplet loss from varying locations within the data tokens. We found that the **specific data token sampled can be chosen arbitrarily and provide extremely consistent results**: with all other factors remaining identical, selecting the data token from indices $\{5, 10, 15, 20, 25\}$ lead to highly similar mAP outcomes with a standard deviation of $\sigma = 0.0006873$.

Anchor/+/-	mAP↑	R1↑	R5↑	R10↑
FFD	**76.6**	**94.3**	**95.3**	**96.0**
DDD	59.0	85.5	89.2	90.4
FFF	57.5	81.2	84.6	86.2
FDD	62.2	80.5	87.3	88.9
DFF	63.6	86.9	89.2	90.0
DFD	60.6	85.2	88.2	90.0
DDF	57.6	83.7	86.9	88.6
FDF	38.2	60.1	63.0	65.0

Table 8. Performance of our method across different Triple Loss Approaches where F=Fusion Token, D=Data Token. For example, *FFD* = anchor (Fusion), positive (Fusion), negative (Data).

7. Conclusion

In this paper, we introduce the Gradual Fusion Transformer (GraFT) for multimodal ReID, a cutting-edge architecture that employs learnable fusion tokens to adeptly capture both modality-specific and object-specific features by guiding self-attention across encoders. Our innovative training paradigm, complemented by an augmented triplet loss, optimizes the ReID feature embedding space, resulting in a more robust model. Through extensive experiments and ablation studies on benchmark datasets RGBNT100 and RGBN300, GraFT not only outperforms existing methods but also offers a new standard in reproducibility, addressing a gap in the current state-of-the-art multimodal ReID methods. To further the utility and adaptability of our approach, we have integrated neural network pruning into GraFT, allowing for a balance between model size and performance, aiming for diverse deployment scenarios. As we advance, our aim is to broaden the applicability of this framework to encompass various application settings and data modalities.

Potential Social Impact. The topic of ReID brings up a nuanced set of trade-offs in terms of societal impacts. On the one hand, ReID has exciting and unique applications in robotics, public safety, criminal investigations, search and rescue, security authentication, and personalized retail experiences. However, potential negative social aspects of ReID and computer vision in general that need to be considered include bias and discrimination from training data, privacy concerns for public spaces, misidentification in criminal cases, and exploitation of data without consent. Ethical implementation under proper regulation, the respect of privacy rights, and transparent deployment of such systems will be crucial for maximizing the benefits of deep learning ReID models in practice.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019. 6

[2] George Barnum, Sabera Talukder, and Yisong Yue. On the benefits of early fusion in multimodal representation learning, 2020. 2

[3] Emrah Basaran, Muhittin Gokmen, and Mustafa E. Kamasak. An efficient framework for visible-infrared cross modality person re-identification, 2019. 2

[4] Qiuyu Chen, Wei Zhang, and Jianping Fan. Cluster-level feature alignment for person re-identification. *arXiv preprint arXiv:2008.06810*, 2020. 6

[5] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15050–15061, June 2023. 2

[6] Ana Cláudia Akemi Matsuki de Faria, Felype de Castro Bastos, José Victor Nogueira Alves da Silva, Vitor Lopes Fabris, Valeska de Sousa Uchoa, Décio Gonçalves de Aguiar Neto, and Claudio Filipi Goncalves dos Santos. Visual question answering: A survey on techniques and common trends in recent literature, 2023. 2

[7] 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. 3

[8] William Falcon et al. Pytorch lightning. https://github.com/PyTorchLightning/pytorch-lightning, 2019. 6

[9] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14750–14759, June 2021. 2

[10] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks, 2019. 7

[11] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. 4

[12] Rohini Goel, Avinash Sharma, and Rajiv Kapoor. Deep learning based thermal object recognition under different illumination conditions. In *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1227–1233, 2021. 2

[13] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[14] Jinbo Guo, Xiaojing Zhang, Zhengyi Liu, and Yuan Wang. Generative and attentive fusion for multi-spectral vehicle re-identification. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pages 1565–1572. IEEE, 2022. 3, 6, 7

[15] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15013–15022, October 2021. 2, 3

[16] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification, 2021. 5

[17] Chaitra Jambigi, Ruchit Rawal, and Anirban Chakraborty. Mmd-reid: A simple but effective solution for visible-thermal person reid. In *British Machine Vision Conference*, 2021. 2, 3

[18] Eleni Kamenou, Jesus Martinez del Rincon, Paul Miller, and Patricia Devlin-Hill. Closing the domain gap for cross-modal visible-infrared vehicle re-identification. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2728–2734. IEEE, 2022. 3, 6

[19] Arnab Karmakar and Deepak Mishra. Pose invariant person re-identification using robust pose-transformation gan, 2021. 2

[20] 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, Apr. 2020. 2, 3, 6

[21] Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. Pytorch distributed: Experiences on accelerating data parallel training. *CoRR*, abs/2006.15704, 2020. 6

[22] Wen Li, Cheng Zou, Meng Wang, Furong Xu, Jianan Zhao, Ruobing Zheng, Yuan Cheng, and Wei Chu. Dc-former: Diverse and compact transformer for person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1415–1423, Jun. 2023. 2, 3

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[24] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1487–1495, Long Beach, CA, USA, June 2019. IEEE. 3, 4, 6

[25] Purvanshi Mehta. Multimodal deep learning, 2020. 2

[26] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention Bottlenecks for Multimodal Fusion. In *NeurIPS*. arXiv, Nov. 2022. 2, 4

[27] Wenjie Pan, Linhan Huang, Jianbao Liang, Lan Hong, and Jianqing Zhu. Progressively hybrid transformer for multi-modal vehicle re-identification. *Sensors*, 23(9), 2023. 3, 7

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[29] Yunjie Peng, Saihui Hou, Chunshui Cao, Xu Liu, Yongzhen Huang, and Zhiqiang He. Deep learning-based occluded person re-identification: A survey, 2022. 2

[30] Shiv Shankar, Laure Thompson, and Madalina Fiterau. Progressive fusion for multimodal integration, 2022. 2

[31] Charu Sharma, Siddhant R. Kapil, and David Chapman. Person re-identification with a locally aware transformer, 2021. 2

[32] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person re-identification, 2022. 2

[33] Lei Tan, Yukang Zhang, Shengmei Shen, Yan Wang, Pingyang Dai, Xianming Lin, Yongjian Wu, and Rongrong Ji. Exploring invariant representation for visible-infrared person re-identification, 2023. 2

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020. 3

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 3

[36] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification, 2018. 2

[37] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2633–2641, 2022. 3

[38] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016. 5

[39] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 3

[40] 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. 1

[41] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C. Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1092–1099. International Joint Conferences on Artificial Intelligence Organization, 7 2018. 2

[42] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications, 2019. 2

[43] Jian'an Zhang, Yuan Yuan, and Qi Wang. Night person re-identification and a benchmark. *IEEE Access*, 7:95496–95504, 2019. 2

[44] Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3529–3537, 2021. 3

[45] Aihua Zheng, Xianpeng Zhu, Zhiqi Ma, Chenglong Li, Jin Tang, and Jixin Ma. Multi-spectral vehicle re-identification with cross-directional consistency network and a high-quality benchmark. *arXiv preprint arXiv:2208.00632*, 2022. 3

[46] 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. 6

[47] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[48] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *CoRR*, abs/1708.04896, 2017. 6

[49] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression, 2017. 7

[50] Zhihui Zhu, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Weishi Zheng, and Xing Sun. Viewpoint-aware loss with angular regularization for person re-identification, 2019. 2