CAMeL: Cross-Modality Adaptive Meta-Learning for Text-Based Person Retrieval

Hang Yu, *Member, IEEE*, Jiahao Wen, and Zhedong Zheng, *Member, IEEE*

***Abstract*—Text-based person retrieval aims to identify specific individuals within an image database using textual descriptions. Due to the high cost of annotation and privacy protection, researchers resort to synthesized data for the paradigm of pretraining and fine-tuning. However, these generated data often exhibit domain biases in both images and textual annotations, which largely compromise the scalability of the pre-trained model. Therefore, we introduce a domain-agnostic pretraining framework based on Cross-modality Adaptive Meta-Learning (CAMeL) to enhance the model generalization capability during pretraining to facilitate the subsequent downstream tasks. In particular, we develop a series of tasks that reflect the diversity and complexity of real-world scenarios, and introduce a dynamic error sample memory unit to memorize the history for errors encountered within multiple tasks. To further ensure multi-task adaptation, we also adopt an adaptive dual-speed update strategy, balancing fast adaptation to new tasks and slow weight updates for historical tasks. Albeit simple, our proposed model not only surpasses existing state-of-the-art methods on real-world benchmarks, including CUHK-PEDES, ICFG-PEDES, and RSTPReid, but also showcases robustness and scalability in handling biased synthetic images and noisy text annotations. Our code is available at https://github.com/Jahawn-Wen/CAMeL-reID**

***Index Terms*—Text-based person retrieval, domain-agnostic, pretraining, cross-modality, meta-learning.**

I. Introduction

TEXT-BASED person retrieval tasks are closely linked with pedestrian re-identification [1], [2] and text-image retrieval [3], [4], aiming to identify specific individuals within an image database using textual queries [5], [6], [7], [8], [9], [10]. With the development of multimodal technologies and the intuitive nature of the text, the demand for matching images using natural language descriptions has been increasingly growing.

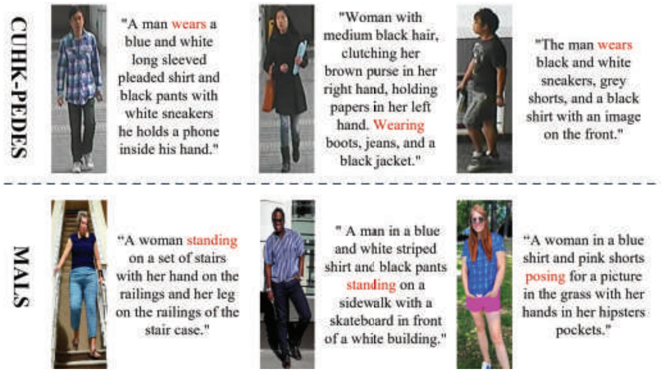

CUHK-PEDES
"A man wears a blue and white long sleeved pleaded shirt and black pants with white sneakers he holds a phone inside his hand."
"Woman with medium black hair, clutching her brown purse in her right hand, holding papers in her left hand. Wearing boots, jeans, and a black jacket."
"The man wears black and white sneakers, grey shorts, and a black shirt with an image on the front."
MALS
"A woman standing on a set of stairs with her hand on the railings and her leg on the railings of the stair case."
" A man in a blue and white striped shirt and black pants standing on a sidewalk with a skateboard in front of a white building."
"A woman in a blue shirt and pink shorts posing for a picture in the grass with her hands in her hipsters pockets."

Fig. 1. Domain biases are observed between the real-world dataset, CUHK-PEDES (top) [5], and the synthetic dataset, MALS (bottom) [14]. The visual domain gap includes facial texture defects, resolution differences, and variations in illumination and color. Text annotations also exhibit bias, with MALS favoring gerunds such as "standing" and "posing," while CUHK-PEDES uses more specific verbs, *e.g.*, "wears."

Compared to traditional image matching methods [11], utilizing natural language descriptions as a query not only simplifies the process of query design, but also enhances the flexibility and intuitiveness of the search. However, due to annotation difficulty and privacy issues, acquiring a large volume of real data for model training is often challenging [12], [13]. In an attempt to address the data scarcity, some researchers [14], [15], [16] have leveraged existing generative models, such as diffusion models [17], [18], to synthesize more training image-text pairs. Although involving generated data facilitates model training, the stylistic biases present in the generated images and textual descriptions often fail to capture the expected representations, thereby limiting the generalization capability of the learned model (see Fig. 1).

Why there are visual domain gaps between real and synthetic data, such as discrepancies in illumination and color, facial texture defects, and resolution? *1) Challenges in Reproducing Real-World Illumination and Color Consistency.* Real-world images are characterized by complex and diverse lighting conditions, including variations in light sources, shadows, and reflections. Additionally, color consistency is influenced by factors such as the color profile of camera, ambient lighting, and post-processing. Generative models often fail to fully capture these nuances, leading to discrepancies in illumination and color. Generated images typically display uniform or unrealistic lighting, and colors often fail to match the subtle variations and natural gradients

Received 4 February 2025; revised 2 April 2025; accepted 23 April 2025. Date of publication 29 April 2025; date of current version 8 May 2025. This work was supported in part by Shanghai Committee of Science and Technology, China, under Grant 23ZR1423500; in part by the National Natural Science Foundation of China under Grant 62302287; in part by Nanjing Municipal Science and Technology Bureau under Grant 202401035; and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515012281. The associate editor coordinating the review of this article and approving it for publication was Prof. Mang Ye. *(Corresponding author: Zhedong Zheng.)*

Hang Yu and Jiahao Wen are with the School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China (e-mail: yuhang@shu.edu.cn; wenjh@shu.edu.cn).

Zhedong Zheng is with the Faculty of Science and Technology and the Institute of Collaborative Innovation, University of Macau, Macau, China (e-mail: zhedongzheng@um.edu.mo).

Digital Object Identifier 10.1109/TIFS.2025.3565392

found in real images. This results in a visual domain gap, where synthetic images appear less realistic and more artificial. For instance, synthetic pedestrian images show flat or uniform lighting, lacking the dynamic shadows and highlights that are typical in real photographs. Colors in the synthetic images also appear washed out or oversaturated, failing to replicate the natural skin tones and environmental colors present in real-world scenes. *2) Inadequate Modeling of High-Frequency Details and Textural Variations.* Generative models, such as GANs and diffusion models, often struggle to accurately capture and reproduce the high-frequency details and textural variations present in real-world images [19]. This is due to inherent model limitations and the complexity of natural textures. The result is that synthetic images usually exhibit artifacts such as blurring, smoothing, or unnatural patterns, which are particularly noticeable in fine-grained regions like skin and hair. These defects can manifest as inconsistencies in facial textures, leading to a clear difference between real and generated images. For example, in some synthetic pedestrian images, the skin appear overly smooth or exhibits unnatural blemishes, while hair lacks the fine details and natural flow seen in real images.

Similarly, we face the textual domain gap in the pedestrian descriptions. To mitigate the image generation failures, such as missing attributes, we typically regenerate captions for synthesized images using the off-the-shelf captioning model. It also introduce the textual biases, which can be attributed to the training data distribution and the way the captioning model is trained. If the dataset used for training the captioning model contains a higher frequency of gerunds, the model will learn to generate captions that reflect this pattern. This is because the training objective is to minimize the loss function, which often leads it to reproduce the most common patterns in the training data. Consequently, the model overfits to the use of gerunds and fail to generalize well to the more varied and contextually rich verb usage found in real-world datasets. For instance, if the training dataset frequently includes images with captions like “A person standing in front of a building” or “A woman posing for a photo,” the captioning model will learn to prefer these gerund forms. In contrast, real-world datasets have more diverse and specific verb usages, such as “A person wears a red jacket” or “A woman smiles at the camera.” Besides, the usage of gerunds in captioning models often simplifies or generalizes actions, making them easier for the model to learn and generate. In a real-world dataset, an image of a person wearing a hat and holding a book is captioned as “A person wears a hat and holds a book.” The verbs “wears” and “holds” provide specific and detailed information about the actions. In a synthetic dataset, the same image could be captioned as “A person standing and holding a book,” where “standing” is a more general description of the state.

Considering the domain gap between the real-world data and generated data, we introduce a domain-agnostic pretraining framework for text-based person retrieval using Cross-modality Adaptive Meta-Learning (CAMeL). In particular, we apply cross-modality adaptive meta-learning strategies to enable the model to identify and adapt to domain-invariant factors across different scenarios, significantly improving its generalization across diverse data environments. Conventional methods that rely solely on image enhancement techniques, such as rotation, often fail to ensure effective feature learning due to the potential for overfitting to specific enhancements, thereby reducing the model’s ability to recognize unenhanced images. To address this, we introduce a dynamic error sample memory unit to store and reuse challenging hard negative samples, leveraging the fast adaptation and transfer learning capabilities of meta-learning. This approach enhances the model’s ability to discern valid combinations of image and text features, leading to more accurate decisions in similar future scenarios. For optimization, we further introduce an adaptive dual-speed update strategy to balance fast adaptability and precise tuning. Fast updates allow the model to rapidly adapt to the basic features of new tasks, while slow updates focus on detailed parameter adjustments, ensuring stability and performance during long-term training. This enables the model to generalize effectively even with limited or incomplete text descriptions, making it well-suited for real-world applications. In summary, our main contributions are as follows:

- We introduce a Cross-modality Adaptive Meta-Learning (CAMeL) to facilitate domain-agnostic pretraining for text-based person retrieval, which mitigates the impact of inherent domain biases in the generated data.
- For multi-task optimization, we further propose a dynamic error sample memory unit and an adaptive dual-speed update strategy to balance the memorization of historical tasks and fast adaptation to new tasks.
- Extensive experiments show that our domain-agnostic pretraining framework via CAMeL has achieved a competitive recall rate on real-world benchmarks, *i.e.*, CUHK-PEDES, ICFG-PEDES, and RSTPReid, surpassing existing methods. Even for the ill-posed text query, *i.e.*, missing several words, the proposed method still yields robust retrieval performance.

The paper is organized as follows. Section II introduces some related works. Section III describes the domain-agnostic pretraining in detail. Section IV discusses the experiment results conducted on some commonly used datasets. Section V concludes the paper and offers suggestions for future work.

II. Related Work

A. Text-to-Image Person Retrieval

Using natural language descriptions for person retrieval is more practical than relying solely on image or attribute queries. Li et al. first propose text-to-image person retrieval and have created the large-scale descriptive dataset CUHK-PEDES [5]. With technological advancements and the diversification of application scenarios, researchers face challenges in accurately understanding and matching the complex relationships between textual descriptions and images. To advance retrieval technology, more complex datasets, such as ICFG-PEDES [20] and RSTPReid [21], have been introduced. The evolution from initial methods focusing on basic feature extraction (*e.g.*, the GNA-RNN model for cross-modality data management [5]) to the current adoption of cross-modal attention mechanisms [22], [23] and deep learning frameworks [24] marks a significant transition. The introduction of attention mechanisms, particularly those utilizing human pose information to locate discriminative regions [25], facilitates multi-granularity feature alignment between images and texts, while Wang et al. [26] focus on the efficiency for the retrieval acceleration. Multi-granularity image-text alignment models by Niu et al. [27] and the adversarial matching approach by Sarafianos et al. [28] further demonstrate the synergistic effects between data at various levels of detail. The VGSG method proposed by He et al. [29] achieves part-level feature alignment through semantic grouping without the need for additional pose alignment tools. Recent studies further refine these approaches. Ergasti et al. [30] highlight the role of visual attributes in text-based person search, while Tan et al. [31] propose a saliency-guided patch transfer method to address occlusion in person re-identification. Zhang et al. [32] enhance cross-modal generalization via middle modality alignment for visible-infrared ReID, and Tan et al. [33] unify data augmentation strategies for cross-spectral ReID. However, the above-mentioned methods do not explicitly deal with the domain gap between the pretraining and the target dataset. In this work, we do not pursue the final performance but focus more on the scalability of the pretrained model for different scenarios.

B. Domain-Agnostic Pretraining

In recent years, with the advancement of machine learning, domain-agnosticism has become a research focus, particularly in reducing a model's dependency on specific domain knowledge and enhancing its generalization capabilities [34], [35], [36]. This direction emphasizes developing models capable of learning data-related distortions during the pretraining phase. By leveraging a range of data sources, these models can acquire general features and patterns in pretraining, enabling rapid adaptation to unseen domains or tasks [37], [38]. Some studies align source domain features with target domain semantics to regularize cross-domain representation learning [39], while the domain-agnostic prompting approach proposed by [40] leverages domain-invariant semantics by aligning visual and textual embeddings. Recent studies have shown that large multilingual models enhance zero-shot multimodal learning across languages [41], while prompt learning research has explored the effect of uninformative class names on generalization [42]. In this study, we explore the use of domain-agnostic pretraining for text-to-image person retrieval and focus on multiple cross-modality tasks to learn domain-invariant features.

C. Meta-Learning

Meta-learning is widely used to improve generalization in cross-modal learning, helping models quickly adapt to new tasks with limited data. Its key strength lies in efficiently learning new tasks using methods like MAML and Reptile, enabling rapid adaptation with minimal data [43], [44], [45]. For instance, Adaptive Uncertainty Learning (AUL) introduces an uncertainty-aware matching mechanism that leverages meta-learning to optimize cross-modal pairs, enhancing generalization on complex datasets [46]. Meta-learning is effective in data-scarce scenarios, enabling rapid adaptation and improved performance [47], [48], [49]. Additionally, Meta-transfer Learning (MTL) addresses domain adaptation by learning meta-feature transformers that enhance adaptability to new domains, especially in unsupervised settings [50]. Meta-learning improves task performance on modality tasks, particularly in handling missing modalities [51]. Moreover, it has been applied to optimize multi-modal alignment, such as image-text alignment, by dynamically adjusting sample weights to prioritize high-quality samples, improving cross-modal alignment [52], [53]. Different from previous works, we introduce meta-learning for domain-agnostic pretraining to minimize the domain gap while facilitating domain-invariant feature learning.

III. Domain-Agnostic Pretraining

Algorithm 1 Cross-Modality Meta-Learning With Adaptive Dual-Speed Updates

```
Require: Initial model parameters $\theta$, two meta-learning rate
    $\epsilon_{slow}$, $\epsilon_{fast}$, update cycle $k$, number of tasks $N$
Ensure: Trained model parameters $\theta'$
Initialize fast parameters $\theta_0 = \theta$
Initialize slow parameters $\theta' = \theta$
for each meta-epoch do
    for each cross-modality task $T_i$ in $\{1, 2, \ldots, N\}$ do
        Initialize $\theta_i = \theta_{i-1}$
        for each training iterations within $T_i$ do
            Sample augmented batch from $T_i$
            Compute loss $\mathcal{L}_{T_i}(\theta_i)$
            # Apply the sequential task-specific updates
            Update task parameters $\theta_i$
        end for
    end for
    # Apply fast update to model parameters
    Aggregate updates: $\theta_0 \leftarrow \theta_0 + \epsilon_{fast}\frac{1}{N}\sum_{i=1}^{N}(\theta_i - \theta_0)$
    if meta-epoch % $k$ == 0 then
        # Apply slow update to model parameters
        Apply slow update: $\theta' \leftarrow \theta' + \epsilon_{slow}(\theta_0 - \theta')$
        Set $\theta_0 = \theta'$
    end if
end for
return $\theta'$
```

Overview. The objective of the domain-agnostic pretraining framework is to mitigate the adverse effects of biases between generated and real data on model training, thus enhancing model performance in text-based person retrieval tasks. In this section, we detail the proposed domain-agnostic pretraining framework for text-based person retrieval, which is depicted in Fig. 2. Firstly, we utilize the generated dataset MALS [14] as our pretraining dataset, designing various tasks such as the Dynamic Illumination Task, Image Blurring Task and Adaptive Memory Task to simulate real-world data complexity. By employing meta-learning strategies, the model is enabled to identify and adapt to key variables across different cross-modality tasks, thereby enhancing its generalization capabilities across diverse data environments. Further, the cross-modal meta-learning strategy uses these varied cross-modality tasks to train the model, allowing fast adaptation to new cross-modality tasks in a short time. In our context, this means the model can learn from various types of data, discerning which information is useful and which could potentially lead to misguidedness. In this way, the model not only learns solutions for specific tasks but also acquires learning skills that can be transferred across different cross-modality tasks. Lastly, to further improve the capability of the model for in-depth exploration of individual tasks, we have introduced an adaptive dual-speed strategy. This allows the model to rapidly assimilate new knowledge while also deeply analyzing and optimizing long-term learning content. Fast updates provide an immediate response to new tasks, while slow updates ensure that more stable and accurate knowledge is refined from these responses, optimizing the long-term performance of the model. The algorithm is summarized in Alg. 1.

A. Stylization Tasks

To enhance generalization capabilities across various application scenarios, a series of tasks have been designed that simulate the complexity and diversity of real-world data. Given the specifics of the dataset generation process, we emphasized image style variations such as lighting, contrast, and occlusion to train the model for diverse visual conditions. In text processing, a range of text transformation techniques have been utilized to simulate different image descriptions. These techniques include synonym replacement, random insertion, random deletion, and random swapping, creating an enriched set of text descriptions. By altering the original text probabilistically to mimic the annotation styles of different experts, the ability to comprehend and match diverse styles and expressions of text descriptions is improved.

Dynamic Illumination Task. Given images from the pretraining dataset MALS, represented as I, we enhance the image dataset I_a by dynamically adjusting the illumination levels. Concurrently, with a certain probability, operations such as random rotation and cropping are applied to alter the visual representation of the images, simulating the appearance of images under different viewing angles and environmental illumination conditions.

Image Blurring Task. During the pretraining phase, we introduced a blurring task by applying Gaussian blur with varying intensities to the images in the set I, generating a blurred image set I_b. This process simulates common image quality issues, such as focus drift and motion blur. The goal of this task is to improve the ability of model to recognize images with partially missing or degraded visual information.

$$I_b(x, y) = I(x, y) \odot G(x, y; \sigma), \tag{1}$$

where $\odot$ is a convolution operation, $G(x, y; \sigma)$ is a Gaussian fuzzy kernel as $G(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$. σ is the standard deviation of the Gaussian kernel, which is used to control the intensity of the ambiguity. This operation not only challenges the ability of model to recognize images with partial loss of visual information but also simulates common image quality issues caused by inaccurate camera focus or motion blur.

Adaptive Memory Task. In text-based task retrieval, data augmentation methods [54], [55] such as rotation, flipping, or color adjustment provide some performance improvements by increasing data diversity. However, these static and predefined methods struggle to adapt to subtle variations in real-world scenarios. To address this issue, we introduce an Adaptive Memory Task that utilizes a dynamic erroneous sample memory unit. The dynamic hard negative sample memory unit takes images I_a and I_b produced by the Dynamic Illumination Task and Image Blurring Task as image inputs, C_a and C_b as textual caption inputs. It employs Mixup [56] and memory bank [57] mechanisms to simulate data variability. Samples from different tasks are dynamically integrated and updated to generate erroneous sample pairs and output real-time adjusted hard negatives. This enables the model to continually learn from and adapt to new and unknown sample features.

Specifically, by linearly interpolating between original input vectors I_a, I_b and their corresponding textual captions C_a, C_b, new vectors $\tilde{I}$ and new captions $\tilde{C}$ are created, merging features from two different data samples to generate hard negative sample pairs. The mixing ratio in the Mixup process is controlled by the parameter λ, which follows a Beta distribution, *i.e.*, $\lambda \sim \mathrm{Beta}(\delta, \delta)$, ensuring a balanced contribution of two sets of tasks during the mixing process. After undergoing two types of augmentations tasks, we obtain two augmented images and captions, I_a, I_b and C_a, C_b. The formulas for the mix can be represented as:

$$\tilde{I} = \lambda \cdot I_\mathrm{a} + (1 - \lambda) \cdot I_\mathrm{b}, \tilde{C} = \lambda \cdot C_\mathrm{a} + (1 - \lambda) \cdot C_\mathrm{b}. \tag{2}$$

Therefore, we could define three typical tasks T_i. T_1 consists of I_a and $\tilde{C}$, while T_2 is composed of I_b and $\tilde{C}$. T_3 contains both $\tilde{I}$ and $\tilde{C}$.

B. Dynamic Error Sample Memory Unit

The memory unit mechanism plays a crucial role in storing and dynamically updating historical information from previous tasks, enabling the model to randomly sample erroneous samples for replaying when learning new tasks. However, this method requires substantial working memory, which is usually infeasible, particularly in text-based pedestrian retrieval tasks where precise matching of images and corresponding text descriptions is necessary. To address this challenge, we focus specifically on identifying and replaying those error samples that pose significant challenges to model performance improvement, known as “hard negatives.” As shown in Fig. 2, when introducing a new batch of hard negatives (*e.g.*, $\tilde{I}$, $\tilde{C}$), if the memory unit is not full, we directly add the embedding of these new samples. If it is full, we first remove some of the old hard negative samples from the unit and then add the new samples in a queue format. This update strategy not only ensures that the data in the memory unit is continuously updated, maintaining its immediacy and relevance to the training process, increasing the difficulty of the image-text matching (ITM) [58] task, but also effectively reduces

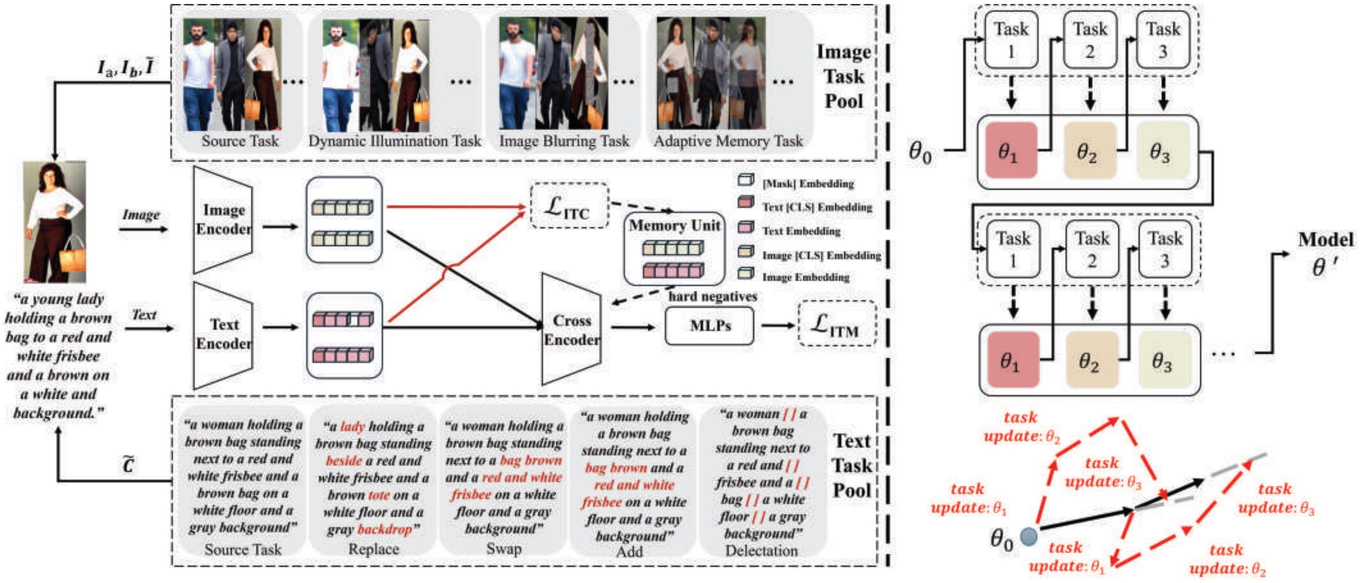

$I_a, I_b, \tilde{I}$
Source Task
Dynamic Illumination Task
Image Blurring Task
Adaptive Memory Task
Image Task Pool
Image
Text
"a young lady holding a brown bag to a red and white frisbee and a brown on a white and background."
Image Encoder
Text Encoder
$\mathcal{L}_{ITC}$
Memory Unit
hard negatives
Cross Encoder
MLPs
$\mathcal{L}_{ITM}$
[Mask] Embedding
Text [CLS] Embedding
Text Embedding
Image [CLS] Embedding
Image Embedding
$\tilde{C}$
"a woman holding a brown bag standing next to a red and white frisbee and a brown bag on a white floor and a gray background"
Source Task
"a lady holding a brown bag standing beside a red and white frisbee and a brown tote on a white floor and a gray backdrop"
Replace
"a woman holding a brown bag standing next to a bag brown and a red and white frisbee on a white floor and a gray background"
Swap
"a woman holding a brown bag standing next to a bag brown and a red and white frisbee on a white floor and a gray background"
Add
"a woman [ ] a brown bag standing next to a red and [ ] frisbee and a [ ] bag [ ] a white floor [ ] a gray background"
Delectation
Text Task Pool
θ_0
Task 1
Task 2
Task 3
θ_1
θ_2
θ_3
Model
θ'
task update: θ_1
task update: θ_2
task update: θ_3
task update: θ_1
task update: θ_2
task update: θ_3
θ_0

Fig. 2. Overview of the proposed domain-agnostic pretraining on the synthetic dataset, *i.e.*, MALS. **(1)** We initially design stylized image tasks involving dynamic illumination, image blurring and adaptive memory, while we apply text augmentation to simulate the real-world natural language inputs. Then augmented image-text pairs are fed into the encoders, and calculate the image-text contrastive loss (ITC) and image-text matching loss (ITM). **(2)** Subsequently, guided by the meta-learning strategy, model parameters are optimized through gradient updates directed by the loss function, adapting to diverse task requirements. The red dashed line represents the task-specific updates, reflecting the model's rapid optimization in specific tasks (lines 6-10 in Alg. 1). The gray dashed line represents the fast update, which helps the model quickly adapt to new tasks by adjusting global parameters (line 14 in Alg. 1). The black line represents the slow update, ensuring gradual convergence through global optimization (line 17 in Alg. 1).

the required storage space by discarding old samples that no longer contribute to learning. Moreover, this approach enhances the ability of model to recognize fine-grained features and classify intra-class variations, effectively simulating the variety of changes encountered in natural environments. Albiet simple, this strategy not only enriches the training batches but also improves the generalization ability of model across different scenarios, significantly enhancing its robustness and consistency in historical tasks.

C. Cross-Modality Meta-Learning

In text-based person retrieval tasks, the model must simultaneously handle data from two distinct modalities: images and text. This requires the model not only to capture information from each individual modality but also to understand and utilize the correlations between them. Therefore, we integrate information from different modalities and employ a cross-modality adaptive meta-learning approach, enabling the model to quickly adapt to new tasks. Meta-learning [43], [44], [59], [60], [61] has been proven to excel in rapid task adaptation, allowing the model to adjust swiftly to new challenges through simple SGD [62] updates. Considering that the model will undergo fine-tuning on new tasks using a gradient-based approach, our learning objective is to enable rapid adaptation to new tasks based on the extraction of internal features, while avoiding over-fitting. This approach ensures rapid progress on new tasks while maintaining generalization capabilities. Specifically, multiple meta-learning epochs are conducted, with each epoch consisting of several tasks. For each cross-modality task T_i, we initialize the task-specific parameters to the last parameters $\theta_i = \theta_{i-1}$. When $i = 0$, we set θ_0 as the initial model parameter θ. Then, θ_i are updated through iterations of gradient descent of the cross-modality task T_i as:

$$\theta_i = \theta_i - \eta \nabla \mathcal{L}_{T_i}(\theta_i), \tag{3}$$

where $\nabla \mathcal{L}_{T_i}(\theta_i)$ is the gradient of the loss function $\mathcal{L}_{T_i}$ with respect to the parameters θ_i for task T_i, and η is the learning rate used for each cross-modality task. Upon completion of all tasks, the model aggregates these updates to obtain fast parameters θ_0 as follows:

$$\theta_0 = \theta_0 + \epsilon_{fast} \frac{1}{N} \sum_{i=1}^{N} (\theta_i - \theta_0), \tag{4}$$

where ϵ_{fast} is the fast meta-learning rate to aggregate task updates, and N is the number of tasks.

D. Adaptive Dual-Speed Update

To effectively implement cross-modality adaptive meta-learning and optimize performance in neural networks, substantial effort is often required to adjust hyperparameters. Particularly in the detailed exploration of independent tasks, to more precisely master the nuances of each task, we have adopted an adaptive dual-speed update strategy. During the fast update phase, the model iterates swiftly by exploring potential search directions. Subsequently, based on the data provided by these fast iterations, the slow update phase integrates this information to optimize and confirm the final direction of model, resulting in a more robust and efficient optimization process. Specifically, at the beginning of model training, we duplicate the model parameters, creating two sets: one for fast updates (denoted as θ_0), and another for slow updates (denoted as θ'), with the initial parameters set as θ. For individual tasks in meta-learning, we use θ_i for regular training optimization. However, after every k iterations of task training, the slow parameters θ' is updated using linear interpolation between θ_0 and θ'. The rule for fast updates θ_0 is presented in Equation 4. Following this, the slow updates are made based on the state after k task iterations. The slow update rule is as follows:

$$\theta' = \theta' \ + \epsilon_{slow}(\theta_0 - \theta'), \tag{5}$$

where ϵ_{slow} is the slow meta-learning rate to control the impact of fast updates on the parameters of the model, confirm the final update direction of model.

IV. Experiment

In this section, we provide a detailed description of the experimental part, which is divided into three main subsections: experimental setup, comparison with competitive methods, and discussion. We begin by introducing three large-scale image-text retrieval datasets, CUHK-PEDES, ICFG-PEDES and RSTPReid, followed by the evaluation metrics, and then detail the implementation specifics. The comparison with competitive methods and thoughts on the experimental study will be discussed in Sections IV-B and IV-C, respectively.

A. Experimental Setup

Datasets. We evaluated our method on three challenging text-to-image person retrieval datasets. **CUHK-PEDES** [5] contains 40,206 images and 80,412 descriptions corresponding to 13,003 identities, with splits of 11,003 for training, 1,000 for validation, and 1,000 for testing. **RSTPReid** [21] is created by compiling MSMT-17 [63] data, includes 20,505 images and 41,010 descriptions for 4,101 individuals, divided into 3,701 identities for training and 400 for testing. **ICFG-PEDES** [20] is also derived from MSMT-17 and consists of 54,522 images and descriptions for 4,102 identities, with 3,102 identities for training and 1,000 for testing.

Evaluation Metrics. To comprehensively evaluate the effectiveness of our proposed model, following previous practices, we adopted Recall@K (where K is 1, 5, 10) as our primary evaluation metric. This metric reflects the ability of the model to successfully identify the target image of the people among the top k most relevant image candidates upon receiving a specific text query. Furthermore, as a supplement to assess the overall retrieval performance of a model, we also introduced the mAP as an evaluation metric. Within this evaluation framework, higher values of Recall@K and mAP indicate superior model performance.

Implementation Details. Our model comprises three encoders: an *image encoder* initialized with 12 layers of SG-Former [83], a *text encoder* initialized with the first 6 layers of BERT [84], and a *cross encoder* initialized with the last 6 layers of BERT. For image augmentation, we employed adaptive techniques such as RandAugment [55], using random horizontal flipping, random erasing, and random cropping, with all images resized to 224 × 224 pixels. For text augmentation, the probabilities for synonym replacement, random insertion, swapping, and deletion were set at 10% for each token. The maximum text sequence length was set to 56, with an embedding dimension of 256. The hyperparameter λ = np.random.beta(δ, δ) with $\delta = 1$. The Baseline (Pretrained) model is trained solely on the synthetic dataset MALS without any fine-tuning on the target dataset, whereas the Baseline (Fine-tuned) model undergoes additional training on the target dataset to better adapt to its distribution. Notably, the augmentation strategies are consistently applied across all experiments to ensure comparability.

Model training proceeds in two phases: pretraining and fine-tuning. During pretraining, we train for 32 epochs on four NVIDIA A6000 GPUs, with a batch size of 80. We use AdamW [85] as the optimizer, with an initial learning rate of 1e-4 using linear decay and a weight decay of 0.01. Following pretraining, the model undergoes fine-tuning on downstream datasets for 30 epochs, with each session lasting around five hours. During the first 10 epochs, the focus is on stylized tasks to enhance data understanding, while the remaining 20 epochs transition into a stable optimization phase, where the learning rate is fixed at 2e-4 for task-specific fine-tuning. A slow update is applied after completing every six cross-modality tasks. Additionally, every 3 epochs, we capture a snapshot of the model's weight parameters and incorporate it into the stochastic weight averaging process [86], balancing generalization and adaptability across cross-modality tasks. We clarify that the model fine-tuning process takes approximately 5 hours. Moreover, the inference time per text query of ours is 5.6ms, validating efficiency comparable to the APTM method.

B. Comparison With Competitive Methods

We compare our method with state-of-the-art text-based person retrieval models on CUHK-PEDES, ICFG-PEDES, and RSTPReid datasets, as detailed in Tables I, II, and III. Our proposed model consistently outperforms existing methods across all three datasets, achieving a competitive recall rate while significantly reducing the number of computational parameters.

On CUHK-PEDES, our method achieves a Recall@1 of 77.24%, Recall@5 of 91.80%, and Recall@10 of 95.16%, with an mAP of 68.32%. In contrast, the second-best method, APTM, records a Recall@1 of 76.53% and an mAP of 66.91%. The consistent improvement, particularly a 0.71% increase in Recall@1, highlights our model's ability to better align fine-grained textual descriptions with image representations. This improvement is attributed to the introduction of cross-modality adaptive meta-learning (CAMeL), which not only enhances generalization across diverse data environments, but also significantly reduces computational complexity through more efficient parameter utilization. Notably, despite the significant reduction in parameters, our model achieves approximately a 5% improvement in every performance metric compared to the TBPS-CLIP model with a similar number of parameters on the CUHK-PEDES dataset, and maintains the same level of improvement across other datasets. This validates that our model delivers excellent performance without increasing computational overhead.

TABLE I

PERFORMANCE COMPARISON ON CUHK-PEDES. BASELINE (PRETRAINED): THE SAME MODEL ARCHITECTURE AS CAMEL, PRE-TRAINED ON THE SAME DATASET, BUT WITHOUT THE INCORPORATION OF ST, CMML, AND ADSU. BASELINE (FINETUNED): THE BASELINE MODEL (PRETRAINED) FURTHER FINE-TUNED ON THE TARGET DATASET. CAMEL (PRETRAINED): THE PRE-TRAINED MODEL APPLIED DIRECTLY TO THE TARGET DATASET WITHOUT FINE-TUNING. CAMEL (FINETUNED): THE CAMEL (PRETRAINED) MODEL FURTHER FINE-TUNED ON THE TARGET DATASET. †: ONLY REPORTS THE NUMBER OF TRAINABLE PARAMETER

Method	#Parameter	R1	R5	R10	mAP
Dual Path [64]	-	44.40	66.26	75.07	-
CMPM+CMPC [65]	-	49.37	-	79.21	-
MIA [27]	-	53.10	75.00	82.90	-
A-GANet [66]	-	53.14	74.03	81.95	-
ViTAA [67]	177M	55.97	75.84	83.52	51.60
IMG-Net [68]	-	56.48	76.89	85.01	-
CMAAM [69]	-	56.68	77.18	84.86	-
HGAN [70]	-	59.00	79.49	86.62	-
NAFS [71]	189M	59.94	79.86	86.70	54.07
DSSL [21]	-	59.98	80.41	87.56	-
MGEL [72]	-	60.27	80.01	86.74	-
SSAN [20]	-	61.37	80.15	86.73	-
NAFS [71]	189M	61.50	81.19	87.51	-
TBPS [73]	43M	61.65	80.98	86.78	-
TIPCB [74]	185M	63.63	82.82	89.01	-
LBUL [52]	-	64.04	82.66	87.22	-
CAIBC [75]	-	64.43	82.87	88.37	-
AXM-Net [76]	-	64.44	80.52	86.77	58.73
SRCF [77]	-	64.88	83.02	88.56	-
LGUR [22]	-	65.25	83.12	89.00	-
CFine [78]	-	69.57	85.93	91.15	-
PLIP-RN50 [79]	-	69.23	85.84	91.16	-
IRRA [3]	194M	73.38	89.93	93.71	66.13
TBPS-CLIP [80]	149M	73.54	88.19	92.35	65.38
RDE [81]	153M	75.94	90.63	94.12	67.56
RaSa [23]	210M	76.51	90.29	94.25	**69.38**
APTM [14]	214M	76.53	90.04	94.15	66.91
WoRA [82]	127M†	76.38	89.72	93.49	67.22
Baseline (Pretrained)	145M	15.97	30.90	40.22	14.77
Baseline (Finetuned)	145M	74.58	88.97	93.63	65.56
CAMeL (Pretrained)	145M	25.26	44.04	52.62	22.52
CAMeL (Finetuned)	145M	**77.24**	**91.80**	**95.16**	68.32

TABLE II

PERFORMANCE COMPARISON ON RSTPREID

Method	#Parameter	R1	R5	R10	mAP
DSSL [21]	-	32.43	55.08	63.19	-
LBUL [52]	-	45.55	68.20	77.85	-
IVT [87]	-	46.70	70.00	78.80	-
CAIBC [75]	-	47.35	69.55	79.00	-
CFine [78]	-	50.55	72.50	81.60	-
IRRA [3]	194M	60.20	81.30	88.20	47.17
TBPS-CLIP [80]	149M	61.96	83.55	88.75	48.26
RDE [81]	153M	65.35	83.95	89.9	50.88
RaSa [23]	210M	66.90	86.50	91.35	52.31
APTM [14]	214M	67.50	85.70	91.45	52.56
WoRA [82]	127M†	66.85	86.45	91.10	52.49
Baseline (Pretrained)	145M	20.90	43.60	54.80	15.50
Baseline (Finetuned)	145M	67.15	86.80	91.95	52.93
CAMeL (Pretrained)	145M	26.60	50.00	60.40	20.51
CAMeL (Finetuned)	145M	**68.50**	**87.40**	**92.70**	**53.61**

TABLE III

PERFORMANCE COMPARISON ON ICFG-PEDES

Method	#Parameter	R1	R5	R10	mAP
Dual Path [64]	-	38.99	59.44	68.41	-
MIA [27]	-	46.49	67.14	75.18	-
ViTAA [67]	-	50.98	68.79	75.78	-
SSAN [20]	-	54.23	72.63	79.53	-
IVT [87]	-	56.04	73.60	80.22	-
LGUR [22]	-	59.02	75.32	81.56	-
CFine [78]	-	60.83	76.55	82.42	-
IRRA [3]	194M	63.46	80.25	85.82	38.06
TBPS-CLIP [80]	149M	65.05	80.34	85.47	39.83
RDE [81]	153M	67.68	82.47	87.36	40.06
RaSa [23]	210M	65.28	80.40	85.12	41.29
APTM [14]	214M	68.51	82.99	87.56	41.22
WoRA [82]	127M†	68.35	83.10	87.53	**42.60**
Baseline (Pretrained)	145M	11.81	25.28	32.99	3.57
Baseline (Finetuned)	145M	66.40	81.49	86.73	39.55
CAMeL (Pretrained)	145M	17.50	33.03	41.04	6.06
CAMeL (Finetuned)	145M	**68.70**	**83.11**	**88.32**	41.58

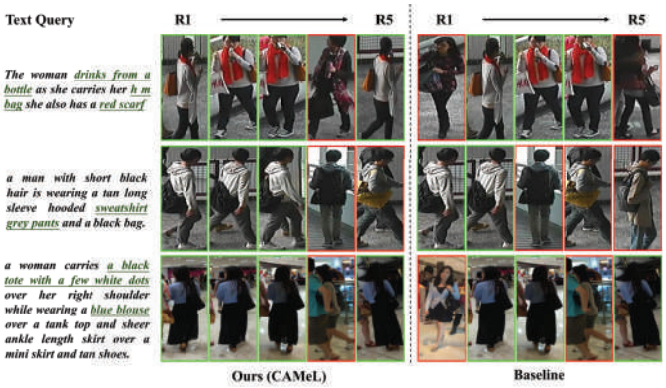

Text Query
R1
R5
R1
R5
The woman drinks from a bottle as she carries her h m bag she also has a red scarf
a man with short black hair is wearing a tan long sleeve hooded sweatshirt grey pants and a black bag.
a woman carries a black tote with a few white dots over her right shoulder while wearing a blue blouse over a tank top and sheer ankle length skirt over a mini skirt and tan shoes.
Ours (CAMeL)
Baseline

Fig. 3. Qualitative comparison of text-to-image retrieval results between Ours (CAMeL) and the Baseline on the benchmark datasets, with results ordered by similarity from highest to lowest, left to right. Correct matches are highlighted with a green frame, while incorrect matches are marked in red. The green-highlighted text emphasizes the details accurately captured by our approach.

For RSTPReid, which focuses on person re-identification across real-world surveillance scenarios with challenging occlusions and viewpoint variations, our method validates an average improvement of 1.25% across key metrics. Importantly, our model's ability to handle difficult conditions in occluded or partially visible scenes is a direct result of the dynamic error sample memory unit, which improves robustness in such scenarios.

On ICFG-PEDES, our method achieves a Recall@1 of 68.70% and an mAP of 41.58%, outperforming the previous best model in all metrics. The robust performance on this dataset further validates the capability of our adaptive dual-speed update (ADSU) strategy to balance fast adaptation with stable, long-term learning, ensuring our model's competitiveness in cross-modal person retrieval tasks. These results validate the effectiveness of our method, not only setting new benchmarks but also demonstrating its robustness and adaptability across diverse real-world scenarios. The proposed approach consistently outperforms state-of-the-art methods,

TABLE IV

Ablation Study on Each Component of CAMeL on Three Benchmark datasets. To Minimize the Impact of Experimental Randomness on Our Research Contributions, We Employed a Method of Averaging the Results From Ten Trials to Present Our Results.Baseline Indicates Pretraining and Fine-Tuning With the Same Dataset and Model Structure, but Not Employing the Proposed CAMeL;ST: Stylization Tasks; ADSU: Adaptive Dual-Speed Update; CMML: Cross-Modal Meta-Learning

No.	Methods	Components			CUHK-PEDES				RSTPReid				ICFG-PEDES			
		ST	ADSU	CMML	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
1	Baseline				74.58	88.97	93.63	65.56	67.15	85.80	91.25	52.93	66.40	81.49	86.73	39.55
2	+ST	✓			75.05	89.86	94.12	66.00	67.50	86.50	91.95	53.14	67.18	81.85	86.90	39.87
3	+ADSU	✓	✓		75.42	90.29	94.19	66.53	67.95	86.90	92.05	52.91	67.64	82.33	87.21	40.56
4	+CMML	✓		✓	76.33	90.58	94.61	67.50	68.30	87.00	92.40	53.43	68.01	82.54	87.59	40.80
5	CAMeL	✓	✓	✓	**77.24**	**91.80**	**95.16**	**68.32**	**68.50**	**87.40**	**92.70**	**53.61**	**68.70**	**83.11**	**88.32**	**41.58**

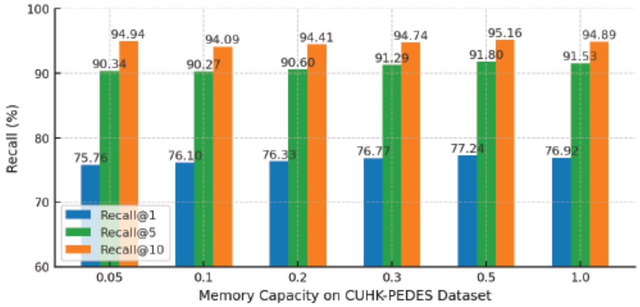


Fig. 4. Ablation Study on the memory capacity in our CAMeL. We apply 5%, 10%, 20%, 30%, 50% and 100% data pairs to pre-train, and then report the fine-tuned performance on CUHK-PEDES dataset. The percentage refers to the current capacity relative to the sample size extracted for dynamic illumination and blurring tasks.

TABLE V

Ablation Study of Tasks Order on the Three Benchmark Datasets. Fixed(Now): Fixed Order. Random: Random Order

CUHK-PEDES	R1	R5	R10	mAP
Fixed (Now)	77.24	**91.80**	95.16	68.32
Random	**77.26** (+0.02)	91.54 (-0.26)	**95.50** (+0.34)	**68.46** (+0.14)
ICFG-PEDES	R1	R5	R10	mAP
Fixed (Now)	**68.70**	83.11	**88.32**	**41.58**
Random	68.46 (-0.24)	**83.27** (+0.16)	88.06 (-0.26)	41.31 (-0.27)
RSTPReid	R1	R5	R10	mAP
Fixed (Now)	**68.50**	**87.40**	**92.70**	**53.61**
Random	68.40 (-0.1)	86.95 (-0.45)	92.55 (-0.15)	53.27 (-0.34)

allowing it to generalize effectively even in challenging settings.

C. Ablation Studies and Further Discussion

Effectiveness of Domain-agnostic Pretraining. We design two sets of ablation studies: the first Baseline (Pretrained) indicates the same model structure with the same pretraining dataset and involves no fine-tuning and directly evaluation on the target tasks; the second acting as CAMeL (Pretrained), which also involves no fine-tuning and directly evaluation on the target tasks. All experiments are carried out under the same dataset and task settings to ensure fairness and comparability of the results. As shown in Table I, models that undergo domain-agnostic pretraining significantly outperform Baseline on all evaluation metrics, underscoring the necessity of domain-agnostic pretraining. Models that do not undergo effective domain-agnostic learning not only perform poorly during the pretraining phase but also negatively impact the results of subsequent fine-tuning. Fig. 3 presents a comparison of retrieval results between our method (CAMeL) and the baseline on the benchmark datasets.

Effectiveness of Stylization Tasks. Considering the importance of pretraining, our experimental approach is based on the Baseline. We have conducted several quantitative experiments to validate the effectiveness of our stylized tasks (ST), as shown in Table IV. By comparing the performance on the Baseline with and without ST, we find that simulating the complexity and diversity of real-world data through ST beneficially influences the learning of model parameters, leading to superior results across three benchmark datasets. Furthermore, to deepen our understanding of the impact of sample library size in adaptive memory tasks, we carry out a series of ablation experiments with sample library sizes ranging from 5% to 100% of the current sample batch. The results indicate that optimal model performance is achieved when the sample library size reaches 50%, as shown in Fig. 4. This is likely because a smaller sample library usually does not adequately cover the diversity of tasks, whereas a larger library introduces redundant information, increasing learning complexity and noise, thereby affecting the model's generalization ability. By optimizing the size of the sample library, we improve the model adaptability to complex scenarios while avoiding overfitting, ensuring the stability and efficiency of the model across various tasks. Moreover, we further investigate the effect of task scheduling in the proposed meta-learning strategy. In our implementation, the model follows a fixed task order: it first performs the Dynamic Illumination Task, then the Image Blurring Task, and finally the Adaptive Memory Task. As shown in Tab. V results across three benchmark datasets show that the task order has negligible impact on final performance. This observation can be explained by the pronounced domain gap between real and synthetic data, as discussed in the Introduction. The three tasks are explicitly designed to simulate key domain characteristics such as lighting conditions, color consistency, and resolution. Their diversity and complementarity reduce sensitivity to task order, ensuring robust generalization across varied data distributions.

Effectiveness of Adaptive Dual-Speed Update. To more precisely master the details of individual tasks, we introduce Adaptive Dual-Speed Update (ADSU). As shown in Table IV, we establish control groups by comparing the sole use of ST with the addition of ADSU across three benchmark datasets, all of which demonstrate superior outcomes. This confirms the effectiveness of the dual-speed update strategy in enhancing the adaptability and stability of the model. Additionally, we explore the impact of various combinations of fast and slow update steps on model performance. We find that setting the update frequency to perform a slow update every six iterations for the three stylized tasks, coupled with a smoothing factor of 0.5, achieves optimal performance. As shown in Tab. VI, ablation studies on the RSTPReid dataset further validate the effectiveness of this configuration. When $k = 6$, the model achieves optimal results in Recall@1, Recall@5, Recall@10 and mAP, reaching 53.61% mAP and 68.50% Recall@1. In contrast, significantly smaller or larger k values disrupt the balance between fast and slow updates, thereby degrading overall performance. This allows the model to rapidly respond to the learning needs of new tasks while maintaining sufficient information accumulation, ensuring that slow updates can proceed after a more comprehensive evaluation of the accumulated learning outcomes.

To further validate the generalization capability of ADSU, we conduct experiments on a cross-modal geo-localization task under diverse weather conditions. This task involves two settings: drone-to-satellite and satellite-to-drone. Specifically, we apply ADSU to a dual-branch neural network and evaluate it on three datasets: University-1652 [88], SUES-200 [89], and CVUSA [90]. In the drone-to-satellite task, the ADSU-enhanced method achieves Recall@1 scores of 66.88%, 55.07%, and 76.26% on these datasets, respectively, outperforming state-of-the-art methods, which score 65.15%, 52.02%, and 75.00%. Notably, under extreme weather conditions such as dark+rain, ADSU demonstrates a 6% improvement. Similar improvements are observed in terms of Average Precision (AP). Consistent performance gains are also observed in the satellite-to-drone task, further underscoring the robustness of ADSU in viewpoint alignment tasks. These results highlight the significant performance improvements that ADSU achieves, particularly under challenging weather conditions, and validate its effectiveness and robustness in cross-modal localization tasks.

Effectiveness of Cross-Modality Meta-Learning. To prevent model overfitting to specific types of tasks during training, which would impede effective learning of image features, we incorporate Cross-Modal Meta-Learning (CMML) as the core of our training strategy. As shown in Table IV, we establish control groups using Stylization Tasks (ST) alone and compare them with experiments that include CMML. On the CUHK-PEDES dataset, we achieve improvements of 1.28%, 0.72%, 0.49%, and 1.50% on Recall@1, 5, 10, and mAP, respectively. Similar results are observed on the RSTPReid and ICFG-PEDES datasets, as shown in Table IV. This highlights the role of CMML in facilitating better alignment between textual and visual data, which is crucial for fine-grained person retrieval. To ensure the fairness and robustness of the training pipeline, we further investigate the impact of pretraining and fine-tuning epochs on model performance. In our setting, the model is pretrained for 32 epochs on the synthetic dataset MALS to learn transferable representations, followed by 30 epochs of fine-tuning on each target dataset to balance generalization and domain-specific adaptation. To validate the effectiveness of this configuration, we conduct additional experiments by extending the fine-tuning epochs to 40, 50, and 60. Results show that prolonged fine-tuning does not lead to performance gains and instead introduces overfitting. Notably, on the CUHK-PEDES dataset, mAP and Recall@1 drop from 68.32% to 66.35% and from 77.24% to 75.48%, respectively. These findings confirm the rationality of the 32 epoch pretraining and 30 epoch fine-tuning configuration, which consistently yields optimal performance across datasets.

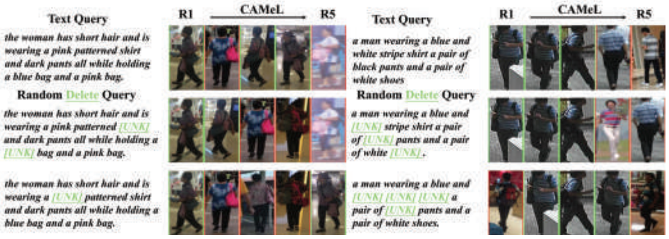

Text Query
R1
CAMeL
R5
the woman has short hair and is wearing a pink patterned shirt and dark pants all while holding a blue bag and a pink bag.
Random Delete Query
the woman has short hair and is wearing a pink patterned [UNK] and dark pants all while holding a [UNK] bag and a pink bag.
the woman has short hair and is wearing a [UNK] patterned shirt and dark pants all while holding a blue bag and a pink bag.
Text Query
R1
CAMeL
R5
a man wearing a blue and white stripe shirt a pair of black pants and a pair of white shoes
Random Delete Query
a man wearing a blue and [UNK] stripe shirt a pair of [UNK] pants and a pair of white [UNK].
a man wearing a blue and [UNK] [UNK] [UNK] a pair of [UNK] pants and a pair of white shoes.

Fig. 5. An example of person retrieval results based on text with randomly masking words is depicted. The retrieved images are arranged from left to right in descending order from R1 to R5. The results validate that increasing the number of deleted words does not impact the precision of our retrieval, confirming the robustness of the CAMeL. The top image-text pair represents the original retrieval result. Green boxes indicate correct matches, while images in red boxes represent incorrect matches.

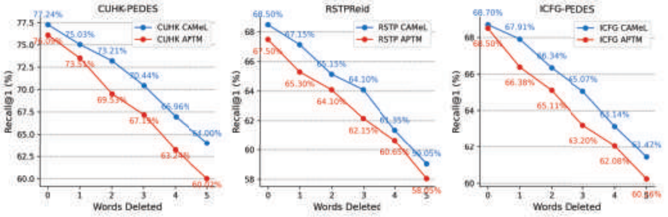
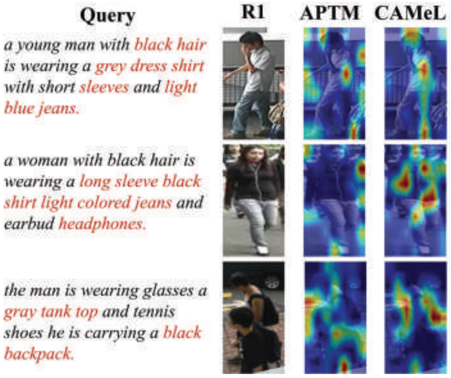


Fig. 6. We assess model performance on the three benchmark datasets by randomly masking words from image annotations and comparing performance before and after deletion. The graph shows Ours (CAMeL) in blue and the APTM* in orange. We could observe that the proposed method is more robust against the ill-posed sentence queries (*e.g.*, missing some words).

Robustness against Ill-formed Text Query. Our pretrained model validates exceptional robustness and generalization capabilities in downstream text-based person retrieval tasks, even with incomplete text queries. After fine-tuning the CAMeL on the CUHK-PEDES dataset, we evaluate its performance by randomly masking 0 to 5 keywords in text queries with the special token [UNK]. As shown in Fig. 5, despite the removal of crucial information such as "clothing," "color," and "style," the model consistently maintains high retrieval accuracy. For example, the model accurately retrieves the image of a woman, even when the word "color" is omitted from the description. Similarly, although the deletion of words like "shirt" and "pants" in the description of the man in the second row leads to one incorrect retrieval, the model still aligned well with the remaining text, illustrating its robustness. Furthermore, we compare our model against the APTM across three datasets using the same word masking technique. As shown in Fig. 6, the gap in Recall@1 between our model and APTM increases significantly, from an initial difference of 1.15% to 3.98% on the CUHK-PEDES dataset,



a young man with black hair is wearing a grey dress shirt with short sleeves and light blue jeans.

a woman with black hair is wearing a long sleeve black shirt light colored jeans and earbud headphones.

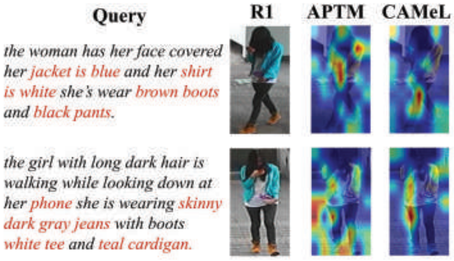
the man is wearing glasses a gray tank top and tennis shoes he is carrying a black backpack.

Fig. 7. Visual comparison of cross attention maps generated by the APTM [14] and Ours (CAMeL) using Grad-CAM [91]. We could observe that the regions highlighted by the proposed methodology exhibit significant alignment with the keywords used in the query sentences, indicating effective matching performance.

with our model also demonstrating more stable performance across the remaining datasets. This uniform enhancement confirms that our model exhibits significantly less fluctuation in Recall@1 with each additional word masking, proving its superior robustness compared to APTM across varied datasets.

Attention Comparison. In Fig. 7, several visual comparisons between the APTM and Ours are presented. Specifically, we utilize the Grad-CAM algorithm [91] to extract attention maps from the models, where each attention map shows the association between the query annotation and the retrieved full-body image of a person. It is evident that the model trained with our method has more focused and consistent attention on each attribute-related word, particularly on attributes such as “grey dress shirt” and “light blue jeans”, where the attention accurately covers the corresponding objects. Furthermore, our training strategy allows the model to generate attention maps that more clearly focus on the correct attributes. For instance, in the “headphones” and “black backpack” attributes, our model exhibits more uniform and reasonable attention distribution compared to the baseline model. In contrast, the APTM model produces irrelevant attention noise in certain image regions, with more scattered attention distribution. Additionally, it is worth noting that even when there are multiple targets present in the image, our training strategy effectively avoids interference from other targets, ensuring that the model can accurately find and focus on the described target. These qualitative results further validate the effectiveness of our proposed training strategy in cross-modal tasks, as it enhances the precise association between text and images, which is crucial for text-based person retrieval tasks.

As shown in Fig 8, the model has effectively captured the pedestrian and their clothing features, aligning well with the textual descriptions. Differences in viewpoint or descriptive emphasis lead to slight shifts in attention, occasionally focusing on background areas. Although the overall performance remains stable, there is still room for improvement-such as further enhancing multi-view robustness, refining text-image alignment, and suppressing irrelevant background in complex scenarios-to achieve more precise feature localization under higher precision and more diverse conditions.



the woman has her face covered her jacket is blue and her shirt is white she's wear brown boots and black pants.

the girl with long dark hair is walking while looking down at her phone she is wearing skinny dark gray jeans with boots white tee and teal cardigan.

Fig. 8. Comparative attention maps with varying queries for the same person.

TABLE VI
Ablation Study of Hyperparameter k on the RSTPReid Datasets

RSTPReid	R1	R5	R10	mAP
k=3	68.20 (-0.30)	86.95 (-0.45)	92.00 (-0.70)	52.76 (-0.85)
k=6 (Now)	**68.50**	**87.40**	**92.70**	**53.61**
k=15	68.50 (-0.00)	86.35 (-1.05)	91.50 (-1.20)	52.54 (-1.07)
k=30	67.75 (-0.75)	87.00 (-0.40)	92.00 (-0.70)	53.16 (-0.45)

Zero-shot Learning. To further validate the generalization capability of the CAMeL, we perform zero-shot experiments on three datasets: CUHK-PEDES [5], RSTPReid [21] and ICFG-PEDES [20]. Without any fine-tuning, we directly test the pre-trained model on the target dataset under identical parameter settings. As shown in Table I, II, and III, our model demonstrates superior generalization ability, achieving improvements in Recall@1, Recall@5, and Recall@10. These results underscore the robustness of the CAMeL strategy employed during pretraining, enabling the model to perform well on unseen datasets without the need for domain-specific fine-tuning. This suggests that the CAMeL approach effectively enhances the model's initial performance, improving its capacity to generalize across different tasks and domains.

Domain Migration. We also conduct domain adaptation experiments to further evaluate the robustness of model. Specifically, the CAMeL fine-tuned on CUHK-PEDES is applied to the ICFG-PEDES dataset. Since both ICFG-PEDES and RSTPReid are derived from the same pretraining dataset, focusing on the CUHK-PEDES to ICFG-PEDES migration ensures a meaningful and challenging domain shift, while also avoiding redundancy from testing two closely related datasets. As shown in Table VII, the model demonstrates stable performance on ICFG-PEDES, validating its ability to adapt to new but related domains. Additionally, we perform a reverse domain migration by fine-tuning the model on RSTPReid and testing it on CUHK-PEDES. This experiment allows us to examine the generalization of model capability when moving from a more controlled surveillance dataset to one with more diverse and varied visual characteristics. The results, as shown in Table VII, reveal that our model maintains competitive performance, illustrating its robustness in adapting to cross-domain variations without extensive re-training. Lastly, we conduct migration experiments between ICFG-PEDES and RSTPReid. Testing these migrations provides additional insights into the model's fine-grained domain adaptation capabilities. The results further support the

TABLE VII

Comparison With Other Pretrained Method. We Adopt CUHK-PEDES (Denoted as C), ICFG-PEDES (Denoted as I) and RSTPReid (Denoted as R) as the Source Domain and the Target Domain in Turn. R@K Is Recall@K (Higher Is better). APTM*: We Re-Implement APTM [14]

	Method	R1	R5	R10
C→I	Dual Path [64]	15.41	29.80	38.19
	MIA [27]	19.35	36.78	46.42
	SSAN [20]	24.72	43.43	53.01
	LGUR [22]	34.25	52.58	60.85
	VGSG [29]	35.85	55.04	63.61
	APTM* [14]	48.57	67.06	74.02
	Ours	**49.18**	**67.58**	**74.63**
I→C	Dual Path [64]	7.63	17.14	23.52
	MIA [27]	10.93	23.77	32.39
	SSAN [20]	16.68	33.84	43.00
	LGUR [22]	25.44	44.48	54.39
	VGSG [29]	27.17	47.77	57.27
	APTM* [14]	46.52	67.53	76.27
	Ours	**70.66**	**87.09**	**91.91**
I→R	APTM*	54.75	77.45	83.90
	Ours	**60.15**	**80.15**	**87.15**
R→I	APTM*	43.11	58.79	65.89
	Ours	**46.34**	**62.88**	**69.81**

generalization of model strength across datasets that are both related and distinct in their own ways.

V. Conclusion

In this paper, we introduce a domain-agnostic pretraining framework that integrates stylized tasks, cross-modality meta-learning, and an adaptive dual-speed strategy to mitigate the negative impact of bias in generated data on model generalization capabilities. By designing stylized tasks, our data better simulates the diversity and complexity of the real world. With the aid of cross-modal meta-learning, we achieve effective integration of information across different tasks, thereby enhancing the generalizability of model. Furthermore, the adaptive dual-speed update strategy delves into the specifics of each task, allowing the model to fast absorb new knowledge while meticulously optimizing the long-term learning process. Experimental results verify that our approach achieves competitive recall rates on the three benchmark datasets. This validates the effectiveness of our proposed pretraining strategy in complex environments. In the future, we plan to explore mechanisms based on erroneous samples to enhance model accuracy. We will probe deeper into the feature representation of erroneous samples. By dynamically identifying and reintegrating these samples, we aim to improve adaptability in challenging scenarios.

References

[1] Z. Zheng and L. Zheng, "Object re-identification: Problems, algorithms and responsible research practice," in *The Boundaries of Data*. Amsterdam, The Netherlands: Amsterdam Univ. Press, 2024.

[2] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," 2016, *arXiv:1604.01850*.

[3] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2787–2797.

[4] H. Shan, Q. Zhang, Z. Liu, G. Zhang, and C. Li, "Beyond two-tower: Attribute guided representation learning for candidate retrieval," in *Proc. ACM Web Conf.*, Apr. 2023, pp. 3173–3181.

[5] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1970–1979.

[6] H. Nguyen, K. Nguyen, S. Sridharan, and C. Fookes, "AG-ReID.v2: Bridging aerial and ground views for person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2896–2908, 2024, doi: 10.1109/TIFS.2024.3353078.

[7] D. Cheng, H. Tai, N. Wang, C. Fang, and X. Gao, "Neighbor consistency and global–local interaction: A novel pseudo-label refinement approach for unsupervised person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 9070–9084, 2024, doi: 10.1109/TIFS.2024.3465037.

[8] A. Lu, C. Li, T. Zha, X. Wang, J. Tang, and B. Luo, "Nighttime person re-identification via collaborative enhancement network with multi-domain learning," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 1305–1319, 2025, doi: 10.1109/TIFS.2025.3527335.

[9] Y. Liu, M. Qi, Y. Zhang, Q. Wu, J. Wu, and S. Zhuang, "Improving consistency of proxy-level contrastive learning for unsupervised person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 6910–6922, 2024, doi: 10.1109/TIFS.2024.3426351.

[10] M. Ye, W. Shen, J. Zhang, Y. Yang, and B. Du, "SecureReID: Privacy-preserving anonymization for person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2840–2853, 2024, doi: 10.1109/TIFS.2024.3356233.

[11] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1575–1590, Apr. 2019, doi: 10.1109/TIP.2018.2878349.

[12] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3754–3762.

[13] G. C. Bertocco, F. Andaló, and A. Rocha, "Unsupervised and self-adaptative techniques for cross-domain person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4419–4434, 2021, doi: 10.1109/TIFS.2021.3107157.

[14] S. Yang, Y. Zhou, Z. Zheng, Y. Wang, L. Zhu, and Y. Wu, "Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 4492–4501.

[15] M. Chu, Z. Zheng, W. Ji, T. Wang, and T.-S. Chua, "Towards natural language-guided drones: GeoText-1652 benchmark with spatial relation matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 1–11.

[16] L. Yao, W. Chen, and Q. Jin, "CapEnrich: Enriching caption semantics for Web images via cross-modal pre-trained knowledge," in *Proc. ACM Web Conf.*, Apr. 2023, pp. 2392–2401.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.

[18] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022, *arXiv:2208.01626*.

[19] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[20] Z. Ding, C. Ding, Z. Shao, and D. Tao, "Semantically self-aligned network for text-to-image part-aware person re-identification," 2021, *arXiv:2107.12666*.

[21] A. Zhu et al., "DSSL: Deep surroundings-person separation learning for text-based person retrieval," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 209–217.

[22] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, and C. Ding, "Learning granularity-unified representations for text-to-image person re-identification," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 5566–5574.

[23] Y. Bai et al., "RaSa: Relation and sensitivity aware representation learning for text-based person search," 2023, *arXiv:2305.13653*.

[24] T. Fujii and S. Tarashima, "BiLMa: Bidirectional local-matching for text-based person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2023, pp. 2778–2782.

[25] S. Wang, H. Li, Z. Wang, and W. Ouyang, "Dynamic position-aware network for fine-grained image recognition," in *Proc. 35th AAAI Conf. Artif. Intell., (AAAI), 33rd Conf. Innov. Appl. Artif. Intell., (IAAI), 11th Symp. Educ. Adv. Artif. Intell., (EAAI)*, 2021, pp. 2791–2799.

[26] X. Wang, Z. Zheng, Y. He, F. Yan, Z. Zeng, and Y. Yang, "Progressive local filter pruning for image retrieval acceleration," *IEEE Trans. Multimedia*, vol. 25, pp. 9597–9607, 2023, doi: 10.1109/TMM.2023.3256092.

[27] K. Niu, Y. Huang, W. Ouyang, and L. Wang, "Improving description-based person re-identification by multi-granularity image-text alignments," *IEEE Trans. Image Process.*, vol. 29, pp. 5542–5556, 2020, doi: 10.1109/TIP.2020.2984883.

[28] N. Sarafianos, X. Xu, and I. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5814–5824.

[29] S. He, H. Luo, W. Jiang, X. Jiang, and H. Ding, "VGSG: Vision-guided semantic-group network for text-based person search," *IEEE Trans. Image Process.*, vol. 33, pp. 163–176, 2023, doi: 10.1109/TIP.2023.3337653.

[30] A. Ergasti, T. Fontanini, C. Ferrari, M. Bertozzi, and A. Prati, "MARS: Paying more attention to visual attributes for text-based person search," 2024, *arXiv:2407.04287*.

[31] L. Tan, J. Xia, W. Liu, P. Dai, Y. Wu, and L. Cao, "Occluded person re-identification via saliency-guided patch transfer," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, 2024, pp. 5070–5078.

[32] Y. Zhang, Y. Yan, Y. Lu, and H. Wang, "Adaptive middle modality alignment learning for visible-infrared person re-identification," *Int. J. Comput. Vis.*, vol. 133, no. 4, pp. 2176–2196, Apr. 2025, doi: 10.1007/s11263-024-02276-4.

[33] L. Tan et al., "RLE: A unified perspective of data augmentation for cross-spectral re-identification," 2024, *arXiv:2411.01225*.

[34] V. Verma, T. Luong, K. Kawaguchi, H. Pham, and Q. V. Le, "Towards domain-agnostic contrastive learning," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10530–10541.

[35] A. Tamkin, V. Liu, R. Lu, D. Fein, C. Schultz, and N. Goodman, "DABS: A domain-agnostic benchmark for self-supervised learning," 2021, *arXiv:2111.12062*.

[36] K. Lee, Y. Zhu, K. Sohn, C.-L. Li, J. Shin, and H. Lee, "I-mix: A domain-agnostic strategy for contrastive representation learning," 2020, *arXiv:2010.08887*.

[37] S. Mishra et al., "Task2Sim: Towards effective pre-training and transfer from synthetic data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9184–9194.

[38] Y. Zhu, H. Shi, Z. Zhang, and S. Tang, "MARIO: Model agnostic recipe for improving OOD generalization of graph contrastive learning," in *Proc. ACM Web Conf.*, May 2024, pp. 300–311.

[39] X. Huo, L. Xie, H. Hu, W. Zhou, H. Li, and Q. Tian, "Domain-agnostic prior for transfer semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7065–7075.

[40] Z. Du, X. Li, F. Li, K. Lu, L. Zhu, and J. Li, "Domain-agnostic mutual prompting for unsupervised domain adaptation," 2024, *arXiv:2403.02899*.

[41] J. Hu et al., "Large multilingual models pivot zero-shot multimodal learning across languages," 2023, *arXiv:2308.12038*.

[42] F. Lv et al., "Rethinking the effect of uninformative class name in prompt learning," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 8345–8354.

[43] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1126–1135.

[44] A. Nichol, J. Achiam, and J. Schulman, "Reptile: A scalable metalearning algorithm," 2018, *arXiv:1803.02999*.

[45] Z. Zhang et al., "Enhancing fairness in meta-learned user modeling via adaptive sampling," in *Proc. ACM Web Conf.*, May 2024, pp. 3241–3252.

[46] S. Li, C. He, X. Xu, F. Shen, Y. Yang, and H. T. Shen, "Adaptive uncertainty-based learning for text-based person retrieval," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, 2024, pp. 3172–3180.

[47] Y. Ma, S. Zhao, W. Wang, Y. Li, and I. King, "Multimodality in meta-learning: A comprehensive survey," *Knowl.-Based Syst.*, vol. 250, Aug. 2022, Art. no. 108976, doi: 10.1016/j.knosys.2022.108976.

[48] Z. Tian, Z. Xie, F. Lin, and Y. Song, "A multi-view meta-learning approach for multi-modal response generation," in *Proc. ACM Web Conf.*, Apr. 2023, pp. 1938–1947.

[49] X. Wang, L. Cao, H. Zhang, L. Feng, Y. Ding, and N. Li, "A meta-learning based stress category detection framework on social media," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2925–2935.

[50] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 403–412.

[51] M. Tran, R. Shah, and Z. Gong, "3FM: Multi-modal meta-learning for federated tasks," 2023, *arXiv:2312.10179*.

[52] Z. Wang et al., "Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 1984–1992.

[53] A. Vettoruzzo, M.-R. Bouguelia, J. Vanschoren, T. Rognvaldsson, and K. Santosh, "Advances and challenges in meta-learning: A technical review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 4763–4779, Jul. 2024, doi: 10.1109/TPAMI.2024.3357847.

[54] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 6382–6388.

[55] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3008–3017.

[56] X. Han, Z. Jiang, N. Liu, and X. Hu, "G-mixup: Graph data augmentation for graph classification," in *Proc. 39th Int. Conf. Mach. Learn.*, vol. 162, 2022, pp. 8230–8248.

[57] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, "MemoryBank: Enhancing large language models with long-term memory," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, 2024, pp. 19724–19731.

[58] Z. Wang, Z. Gao, X. Xu, Y. Luo, Y. Yang, and H. T. Shen, "Point to rectangle matching for image text retrieval," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4977–4986.

[59] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," 2018, *arXiv:1803.02999*.

[60] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–11.

[61] Z. Wu, X. Wang, J. Gonzalez, T. Goldstein, and L. Davis, "ACE: Adapting to changing environments for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2121–2130.

[62] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, Sep. 1951, doi: 10.1214/aoms/1177729586.

[63] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[64] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–23, May 2020, doi: 10.1145/3383184.

[65] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comp. Vis.*, 2018, pp. 686–701.

[66] J. Liu, Z.-J. Zha, R. Hong, M. Wang, and Y. Zhang, "Deep adversarial graph attention convolution network for text-based person search," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 665–673.

[67] Z. Wang, Z. Fang, J. Wang, and Y. Yang, "ViTAA: Visual-textual attributes alignment in person search by natural language," in *Proc. Eur. Conf. Comp. Vis.* Cham, Switzerland: Springer, 2020, pp. 402–420.

[68] Z. Wang, A. Zhu, Z. Zheng, J. Jin, Z. Xue, and G. Hua, "IMG-net: Inner-cross-modal attentional multigranular network for description-based person re-identification," *J. Electron. Imag.*, vol. 29, no. 4, Aug. 2020, Art. no. 043028, doi: 10.1117/1.jei.29.4.043028.

[69] S. Aggarwal, R. V. Babu, and A. Chakraborty, "Text-based person search via attribute-aided matching," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2617–2625.

[70] K. Zheng, W. Liu, J. Liu, Z.-J. Zha, and T. Mei, "Hierarchical Gumbel attention network for text-based person search," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3441–3449.

[71] C. Gao et al., "Contextual non-local alignment over full-scale representation for text-based person search," 2021, *arXiv:2101.03036*.

[72] C. Wang, Z. Luo, Y. Lin, and S. Li, "Text-based person search via multi-granularity embedding learning," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1068–1074.

[73] X. Han, S. He, L. Zhang, and T. Xiang, "Text-based person search with limited data," 2021, *arXiv:2110.10807*.

[74] Y. Chen, G. Zhang, Y. Lu, Z. Wang, and Y. Zheng, "TIPCB: A simple but effective part-based convolutional baseline for text-based person search," *Neurocomputing*, vol. 494, pp. 171–181, Jul. 2022, doi: 10.1016/j.neucom.2022.04.081.

[75] Z. Wang et al., "CAIBC: Capturing all-round information beyond color for text-based person retrieval," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 5314–5322.

[76] A. Farooq, M. Awais, J. Kittler, and S. S. Khalid, "AXM-Net: Implicit cross-modal feature alignment for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 4477–4485.

[77] W. Suo et al., "A simple and robust correlation filtering method for text-based person search," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 726–742.

[78] S. Yan, N. Dong, L. Zhang, and J. Tang, "CLIP-driven fine-grained text-image person re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 6032–6046, 2023, doi: 10.1109/TIP.2023.3327924.

[79] J. Zuo et al., "PLIP: Language-image pre-training for person representation learning," 2023, *arXiv:2305.08386*.

[80] M. Cao, Y. Bai, Z. Zeng, M. Ye, and M. Zhang, "An empirical study of clip for text-based person search," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, 2024, pp. 465–473.

[81] Y. Qin, Y. Chen, D. Peng, X. Peng, J. Tianyi Zhou, and P. Hu, "Noisy-correspondence learning for text-to-image person re-identification," 2023, *arXiv:2308.09911*.

[82] J. Sun, H. Fei, G. Ding, and Z. Zheng, "From data deluge to data curation: A filtering-WoRA paradigm for efficient text-based person search," in *Proc. ACM Web Conf.*, Apr. 2025, pp. 2341–2351.

[83] S. Ren, X. Yang, S. Liu, and X. Wang, "SG-former: Self-guided transformer with evolving token reallocation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 5980–5991.

[84] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[85] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[86] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. Gordon Wilson, "Averaging weights leads to wider optima and better generalization," 2018, *arXiv:1803.05407*.

[87] X. Shu et al., "See finer, see more: Implicit modality alignment for text-based person retrieval," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2022, pp. 624–641.

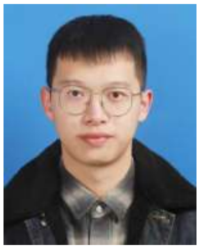
[88] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1395–1403.

[89] R. Zhu, L. Yin, M. Yang, F. Wu, Y. Yang, and W. Hu, "SUES-200: A multi-height multi-scene cross-view image benchmark across drone and satellite," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4825–4839, Sep. 2023, doi: 10.1109/TCSVT.2023.3249204.

[90] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4132–4140.

[91] J. Gildenblat. (2021). *Contributors, PyTorch Library for Cam Methods*. [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam

Hang Yu (Member, IEEE) received the Ph.D. degree from the University of Technology Sydney, Australia, in 2020. He is currently a Professor with the School of Computer Engineering and Science, Shanghai University, China. He has authored or co-authored more than 60 publications and his publications have appeared in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, and IEEE TRANSACTIONS ON FUZZY SYSTEMS. His research interests include streaming data mining, concept drift, and fuzzy systems. He also regularly serves as a program committee member for numerous national and international conferences. He was awarded the Outstanding Academic Leader of Shanghai.



Jiahao Wen is currently pursuing the Ph.D. degree in computer science with the School of Computer Engineering and Science, Shanghai University, Shanghai, China. His research interests include image retrieval, graph neural networks, and meta-learning.

Zhedong Zheng (Member, IEEE) received the B.S. degree from Fudan University, Shanghai, China, in 2016, and the Ph.D. degree from the University of Technology Sydney, Sydney, NSW, Australia, in 2021. He was a Research Fellow with the School of Computing, National University of Singapore, Singapore. He is currently an Assistant Professor with the University of Macau, Macau, China. He received the IEEE Circuits and Systems Society Outstanding Young Author Award in 2021. He also serves as an Area Chair for ACM MM 2024. He has organized a special session on reliable retrieval at ICME 2022, two workshops at ACM MM 2023, and one workshop at ACM ICMR 2024. Besides, he is invited as a Keynote Speaker at CVPR 2020 and CVPR 2021, and a Tutorial Speaker at ACM MM 2022.