

A Theory-Inspired Framework for Few-Shot Cross-Modal Sketch Person Re-Identification

Yunpeng Gong^{1,2}, Yongjie Hou³, Jiangming Shi⁴, Kim Long Diep^{1,2}, Min Jiang^{1,2*}

¹ Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University

²Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan,
Ministry of Culture and Tourism, Xiamen University

³ School of Electronic Science and Engineering, Xiamen University

⁴ Institute of Artificial Intelligence, Xiamen University

gongyunpeng@stu.xmu.edu.cn or fmonkey625@gmail.com, {23120231150268,31520240158147}@stu.xmu.edu.cn,
jiangming.shi@outlook.com, minjiang@xmu.edu.cn

Abstract

Sketch-based person re-identification aims to match hand-drawn sketches with RGB surveillance images, but remains challenging due to severe modality gaps and limited labeled data. To address this, we propose KTCAA, a theoretically inspired framework for few-shot cross-modal generalization. Drawing on generalization bounds, we identify two key factors affecting target risk: (1) domain discrepancy, reflecting the alignment difficulty between source and target distributions; and (2) perturbation invariance, measuring the model’s robustness to modality shifts. Accordingly, we design: (1) Alignment Augmentation (AA), which applies localized sketch-style transformations to simulate target distributions and guide progressive alignment; and (2) Knowledge Transfer Catalyst (KTC), which enhances perturbation invariance by introducing worst-case modality perturbations and enforcing consistency. These modules are jointly optimized within a meta-learning paradigm that transfers alignment knowledge from data-abundant RGB domains to sketch scenarios. Experiments on multiple benchmarks show that KTCAA achieves state-of-the-art performance, particularly under data-scarce conditions.

Code — https://github.com/finger-monkey/REID_KTCAA

Introduction

Person re-Identification (ReID) (Zhong et al. 2019, 2017; Liu, Zhang, and Yang 2020; Sun et al. 2024b; Wang, Gong, and Yan 2024; Gong, Huang, and Chen 2022; Gong et al. 2024a) aims to match individuals across different images or camera views, playing a key role in surveillance and law enforcement. However, in situations where photographic evidence is unavailable, Sketch ReID becomes essential. It focuses on matching hand-drawn sketches, often provided by eyewitnesses, with RGB surveillance images. This task is inherently challenging due to the large modality gap—sketches lack color, texture, and fine-grained detail,

and they vary significantly in drawing style. Moreover, collecting large-scale annotated sketch datasets is costly, making generalization in real-world scenarios even more difficult. Existing sketch ReID methods (Pang et al. 2018; Gong, Huang, and Chen 2021; Ge, Li, and Zhang 2024; Chen et al. 2024) primarily focus on aligning cross-modal features and learning modality-invariant embeddings. While effective in controlled benchmarks, they often rely on abundant labeled sketch data and struggle to generalize under limited supervision. This label scarcity, coupled with high inter-domain variation, calls for a learning framework that can transfer foundational visual knowledge to sketch domains using only a few examples.

To this end, we propose KTCAA, a meta-learning-based framework designed for few-shot cross-modal ReID. Inspired by recent advances in meta-learning (Sung et al. 2018), our method transfers knowledge from large-scale single-modal RGB data to address data-scarce cross-modal scenarios. The training process simulates low-shot tasks through meta-training and meta-testing phases, enabling the model to acquire transferable knowledge and adapt quickly to new domains with minimal sketch data. Critically, unlike previous works (Pang et al. 2018; Gong, Huang, and Chen 2021; Liu et al. 2024b; Lin et al. 2023; Zhu et al. 2022; Ge, Li, and Zhang 2024; Chen et al. 2024), KTCAA is not merely a framework design, but is motivated by theoretical inspiration. We first revisit cross-modal transfer from the perspective of domain adaptation theory. By establishing a generalization error bound, we identify two necessary conditions for effective cross-modal transfer: domain discrepancy and perturbation invariance. These two conditions inform the design of our method.

To satisfy these conditions, we propose two theoretically inspired components. First, the Alignment Augmentation (AA) module reduces the distribution divergence between RGB and sketch modalities by performing local sketch-style transformations on RGB inputs. This localized augmentation simulates domain shifts while preserving identity structure. Second, the Knowledge Transfer Catalyst (KTC) module improves the model’s robustness to modality shifts

*Corresponding author: Min Jiang.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

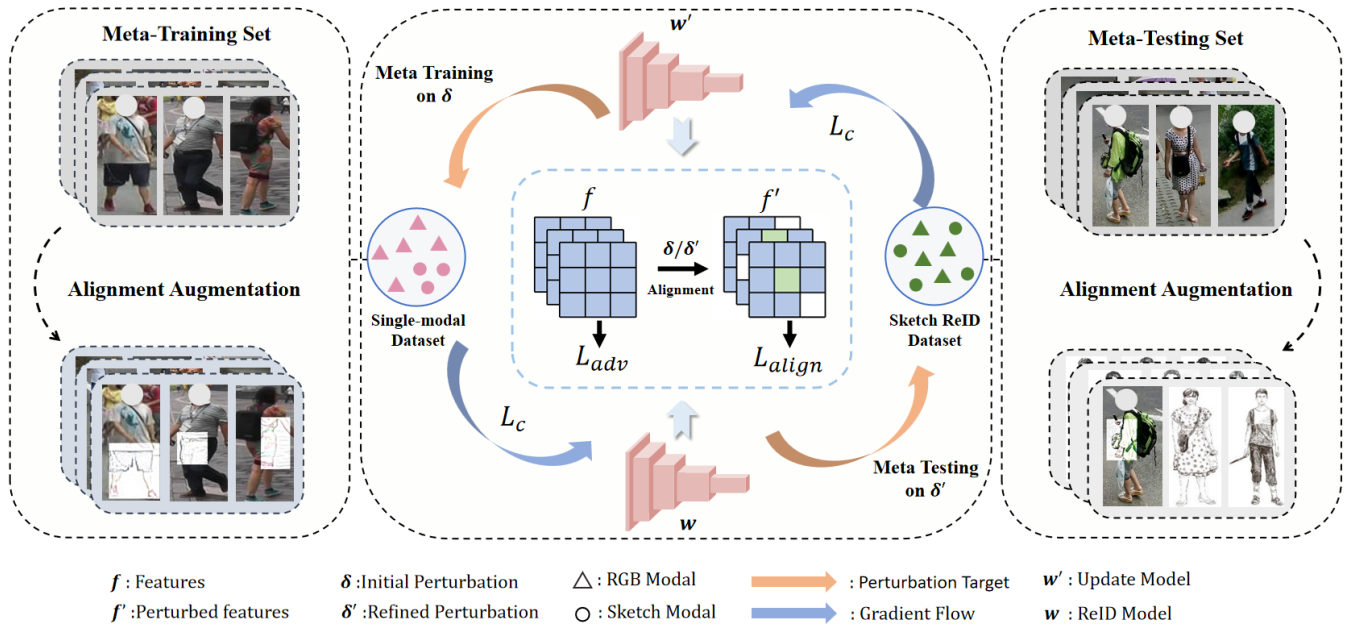


Figure 1: Pipeline of the proposed KTCAA framework, which operates under a meta-learning paradigm. During the meta-training phase, a single-modal RGB dataset is used. The Alignment Augmentation (AA) module applies localized sketch-style transformations to simulate target domain characteristics, modeling modality discrepancies at the image level. This guides the model to progressively align source and target distributions while preserving fine-grained semantics. The Knowledge Transfer Catalyst (KTC) module introduces adversarial perturbations to simulate cross-modal uncertainty and is jointly optimized through meta-learning. The alignment loss L_{align} between features before and after perturbation, along with the adversarial classification loss L_{adv} , enhances the model’s robustness against detail blur and modality shifts. Additionally, the contrastive loss L_c is jointly optimized with these regularization terms to enhance cross-modal representation learning. During the meta-testing phase, the base model parameters w are frozen, and the updated model w' is fine-tuned for few-shot sketch-based Re-ID, leveraging cross-modal knowledge for improved generalization under domain shifts.

by introducing worst-case perturbations that simulate cross-modal variations. These components are jointly optimized under a meta-learning scheme that promotes robust, domain-adaptive feature representations.

An overview of KTCAA is shown in Fig.1. The AA module guides the model to learn shared visual cues across distributions through partial modality transformation. During meta-training, the local differences introduced by AA are used to generate adversarial gradients in KTC, which intentionally amplify modality-induced feature discrepancies. This process transforms the destabilizing nature of adversarial perturbations into a constructive mechanism for learning transferable representations. By aligning representations across modality variations, the model learns modality-consistent embeddings that generalize well to unseen sketches. In KTCAA, knowledge transfer is realized through adversarial training with shared perturbation priors across the meta-learning stages. Specifically, during meta-training on the RGB domain, the model is encouraged to learn modality perturbation-invariant representations via the KTC module. These adversarial priors, rather than being specific to a particular domain, encode general robustness to modality shifts. During meta-testing, this generalizable robustness is transferred to sketch-based tasks, enabling the model to adapt effectively under few-shot conditions. This

mechanism allows KTCAA to transfer cross-modal knowledge implicitly.

The main contributions of this paper are summarized as follows:

- This paper presents a heuristic analysis to establish a generalization bound for cross-modal transfer, revealing domain discrepancy and perturbation invariance as two key factors affecting performance. Based on this insight, we propose a principled framework that incorporates two dedicated modules—Alignment Augmentation and Knowledge Transfer Catalyst—designed to address these challenges.
- We propose KTCAA, a progressive meta-learning framework that effectively adapts from RGB to sketch domains with limited data. Experiments demonstrate that our method achieves superior performance and generalization compared to state-of-the-art approaches.

Related Works

Cross-modal Person Re-identification

With the development of deep learning (Pu et al. 2025; Gu et al. 2024; Sun et al. 2024a; Yang, Huang, and Peng 2024; Zheng et al. 2025; Yang et al. 2022c; Huang et al. 2025), cross-modal recognition techniques have made significant progress, offering new possibilities for addressing identity

recognition challenges in complex surveillance and law enforcement scenarios. Cross-modal ReID is a critical task in security (Gong et al. 2024b; Gu et al. 2025) and surveillance applications, aiming to match pedestrian images captured across different devices and modalities. In scenarios where RGB images are unavailable, such as in low-light or night-time conditions, alternative modalities like infrared are essential. Traditional single-modal ReID (Pang et al. 2025b,a; Liu et al. 2024a; Lu et al. 2025; Sun and Zheng 2019) focuses exclusively on RGB images, whereas cross-modal ReID (Li et al. 2025b; Yang, Chen, and Ye 2024; Yang et al. 2022b, 2023; Li et al. 2025a,c; Zhang et al. 2024; Yang et al. 2022a) addresses the challenges of aligning features across multiple modalities.

Sketch ReID (Gong, Huang, and Chen 2021; Pang et al. 2018; Zhu et al. 2022; Ge, Li, and Zhang 2024; Chen et al. 2024; Lin et al. 2023) represents a unique and particularly challenging cross-modal task due to the substantial semantic gap between sketches and real-world images. Unlike RGB or infrared images, sketches lack critical details such as color and texture, and they are often subject to stylistic variations introduced by different artists. This introduces significant inter-modality discrepancies. To address these challenges, prior work has proposed various solutions. Pang et al. (Pang et al. 2018) applied cross-domain adversarial learning to identify domain-invariant features between sketches and RGB images. Subsequent research explored semantic feature alignment (Chen et al. 2023) to reduce inconsistencies between these modalities. Chen et al. (Liu et al. 2024b) introduced Differentiable Auxiliary Learning (DAL), which generates auxiliary modalities using deep neural networks to bridge the gap between sketches and RGB images. However, the scarcity of large annotated sketch datasets remains a major limitation, often leading to overfitting and poor generalization across datasets. Additionally, SS-reID (Lin et al. 2023) have explored the use of pretrained multimodal models like CLIP (Radford et al. 2021) to aid feature alignment across modalities. While this approach can improve alignment in certain scenarios, it typically relies on additional attribute labels within datasets, which limits its applicability in more generalized settings. Furthermore, the substantial modality gap may restrict the effectiveness of CLIP in achieving robust alignment.

Meta-Learning

Meta-learning has been developed to address the challenge of adapting to new tasks with limited training samples, thereby enhancing generalization across diverse scenarios (Zhang et al. 2023; Guo et al. 2020). Existing meta-learning approaches are typically categorized into three groups: metric-based methods (Sung et al. 2018), model-based methods (Santoro et al. 2016), and optimization-based methods (Finn, Abbeel, and Levine 2017). Among these, MAML (Model-Agnostic Meta-Learning) (Finn, Abbeel, and Levine 2017), an optimization-based approach, has gained prominence for its ability to learn well-initialized weights that can quickly adapt to new tasks by simulating the learning process on meta-test tasks.

Building on the principles of meta-learning, we aim to

address the challenges of cross-modal ReID by leveraging abundant single-modal data to inject shared knowledge into few-shot cross-modal tasks. Our goal is to enhance the generalization of sketch-based re-identification tasks, particularly in unseen domains. By progressively refining shared feature representations and reducing modality discrepancies across meta-training and meta-testing, we enable more efficient cross-modal transfer without requiring extensive supervision from the target modality.

Theoretical Motivation

Sketch ReID faces the fundamental challenge of generalizing from the RGB modality (source domain) to the sketch modality (target domain) under small-sample conditions. To address this, we investigate the problem from a generalization perspective and, drawing inspiration from domain adaptation theory, arrive at an extended formulation that highlights two key factors: domain discrepancy and perturbation sensitivity. This formulation provides theoretical inspiration for our framework design.

Cross-Modal Generalization Bound

Following domain adaptation theory (Ben-David et al. 2010), the expected target domain risk can be decomposed into three components: the expected source domain risk, a distribution discrepancy term, and an irreducible residual:

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda^*, \quad (1)$$

where $\epsilon_T(h)$ and $\epsilon_S(h)$ denote the expected risks of a hypothesis h on the target and source domains, respectively, $d_{\mathcal{H}\Delta\mathcal{H}}$ measures the discrepancy between the two domains, and λ^* represents the minimum joint error of the ideal hypothesis, typically considered task-specific and fixed.

We assume that the feature extractor f is L -Lipschitz continuous and exhibits local robustness to input perturbations δ , such that:

$$\mathbb{E}_{x \sim \mathcal{D}_S} \|f(x + \delta) - f(x)\| \leq \gamma. \quad (2)$$

Under this assumption, the source risk can be approximated by the empirical risk plus a perturbation-sensitive regularization term:

$$\epsilon_S(h) \approx \epsilon_S(f) \leq \hat{\epsilon}_S(f) + L \cdot \gamma, \quad (3)$$

where $\hat{\epsilon}_S(f)$ denotes the empirical source risk, and $L \cdot \gamma$ reflects the worst-case impact of input perturbations.

Inspired by robustness principles such as Virtual Adversarial Training (VAT) (Miyato et al. 2018) and adversarial data augmentation (Volpi et al. 2018), we introduce the following extended heuristic generalization bound to capture the key factors influencing cross-modal transfer:

$$\epsilon_T(h) \lesssim \hat{\epsilon}_S(f) + L \cdot \gamma + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}} + \lambda^*. \quad (4)$$

The term $L \cdot \gamma$ accounts for the model’s sensitivity to input variations and provides a useful regularization perspective inspired by adversarial robustness. Although it is not part of the original bound proposed by Ben-David et al. (Ben-David et al. 2010), we adopt it as a conceptual objective to guide the design of our framework.

In this formulation, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ quantifies the discrepancy between source and target domains, reflecting how differently classifiers behave across modalities. To reduce this, the AA module locally transforms RGB images into sketch-like variants, aligning distributions at the input level and narrowing the domain gap.

Meanwhile, γ captures the model’s sensitivity to input perturbations such as cross-modal shifts. Unlike $d_{\mathcal{H}\Delta\mathcal{H}}$, which focuses on data-level differences, γ characterizes the stability of the feature extractor itself. To suppress it, the KTC module applies worst-case perturbations and enforces consistency through adversarial and alignment losses, thereby promoting robust and transferable representations.

AA: Reducing Domain Discrepancy

To reduce the domain divergence term $d_{\mathcal{H}\Delta\mathcal{H}}$, we introduce the Alignment Augmentation (AA) module, which simulates sketch-style variations by locally replacing regions in RGB images with their sketch-transformed counterparts. This localized perturbation preserves global structure while introducing target-modality characteristics.

Formally, given an input image x and its sketch-transformed version $\Delta_m(x)$, we define:

$$\delta_{\text{global}} = \Delta_m(x) - x, \quad \delta_{\text{local}} = M \odot (\Delta_m(x) - x), \quad (5)$$

where $M \in \{0, 1\}^{H \times W}$ is a binary mask selecting the perturbed region, and \odot denotes element-wise multiplication.

Assuming the feature extractor f is L -Lipschitz, we have:

$$\|f(x + \delta) - f(x)\| \leq L \cdot \|\delta\|. \quad (6)$$

Since $\|\delta_{\text{local}}\| < \|\delta_{\text{global}}\|$, the resulting feature perturbation is smaller under local augmentation. This allows the model to preserve semantic structure while introducing modality-specific features more smoothly, thereby facilitating cross-domain alignment and effectively reducing $d_{\mathcal{H}\Delta\mathcal{H}}$.

KTC: Minimizing Perturbation Sensitivity

To suppress the perturbation response γ , we adopt a robustness-inspired training objective based on worst-case input shifts. Inspired by VAT, we approximate the expected perturbation sensitivity via an adversarial upper bound:

$$\gamma = \mathbb{E}_x \|f(x + \delta) - f(x)\| \lesssim \max_{\|\eta\| \leq \epsilon} \|f(x + \eta) - f(x)\|. \quad (7)$$

Here, the left-hand side represents the model’s average sensitivity to random input perturbations δ , while the right-hand side approximates the worst-case response under bounded adversarial perturbations η satisfying $\|\eta\| \leq \epsilon$. This upper bound serves as a tractable surrogate that captures the model’s robustness to modality shifts. To implement this, the Knowledge Transfer Catalyst (KTC) module generates adversarial sketch-style perturbations and enforces feature consistency via the alignment loss:

$$\mathcal{L}_{\text{align}} = \|f(x + \eta) - f(x)\|_2^2. \quad (8)$$

This encourages the model to learn perturbation-invariant representations and effectively reduces γ , supporting better cross-modal generalization.

Methodology

Motivated by the theoretical analysis of the generalization bound in cross-modal learning, we identify two key factors for robust transfer: domain discrepancy and perturbation stability. These insights guide the design of our proposed framework, KTCAA, which adopts a dual-path architecture based on ResNet50. It extracts modality-specific features for RGB and sketch images through independent shallow branches, while sharing upper fully connected layers for unified representation. The framework consists of two theory-inspired modules: Alignment Augmentation (AA), which reduces domain discrepancy and facilitates progressive alignment through localized modality transformations that simulate target sketch-style characteristics; and Knowledge Transfer Catalyst (KTC), which improves stability under latent modality variations by enforcing consistency under adversarial modality perturbations. These two modules are integrated into a meta-learning strategy designed to improve adaptation in cross-domain few-shot scenarios through joint optimization.

Meta-Training

Alignment Augmentation. The process of alignment augmentation can be represented as follows:

$$x_i^{\text{sketch}} = t(x_i^{\text{rgb}}), \quad (9)$$

where $t(\cdot)$ denotes the function that converts an RGB image into a sketch image. Furthermore, if it is necessary to select a local region at a random position in the image for sketching, this process can be expressed as:

$$x_i^{\text{ls}} = LS(x_i^{\text{rgb}}, x_i^{\text{sketch}}, \text{rect}), \quad (10)$$

where rect represents the randomly selected rectangular area, and $LS(\cdot)$ indicates the operation of replacing the local region in the RGB image with the corresponding sketch. More details can be found in the Appendix.

We use InfoNCE loss as the contrastive loss, which is defined as:

$$L_c = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i^+)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}, \quad (11)$$

where \mathbf{z}_i denotes the feature representation of the original RGB image, and \mathbf{z}_i^+ represents the feature representation of either the alignment augmentation sample or a sketch sample of the same identity. Here, \mathbf{z}_j refers to the representation of negative samples with different identities from \mathbf{z}_i , $\text{sim}(\cdot)$ is a similarity function, τ is a temperature parameter that controls the smoothness of the distribution (set to $\tau = 0.1$), and N is the batch size. This loss function encourages the model to learn discriminative features by maximizing the similarity between positive pairs (\mathbf{z}_i and \mathbf{z}_i^+) while minimizing the similarity between negative pairs (\mathbf{z}_i and \mathbf{z}_j).

Knowledge Transfer Catalyst. The goal of the Knowledge Transfer Catalyst is to optimize knowledge transfer by learning universal perturbations (Moosavi-Dezfooli et al. 2017; Gong et al. 2024b). By maximizing potential differences between modalities, these perturbations encourage the

model to learn robust feature representations. In the meta-learning and meta-training phases, the information interactions facilitated by these universal perturbations expose the model to diverse cross-modal variations, thereby improving the stability of feature learning and enhancing the effectiveness of knowledge transfer.

We design triplet loss to distort the pairwise relationships between pedestrian identities. For a specific batch of n samples that includes both original and alignment augmentation images, we utilize the model to extract and perturb their features. The temporary perturbation η is iteratively updated using Momentum Inertia Stochastic Gradient Descent (MI-SGD), as shown below:

$$L_{RGB}(f_{RGB}^{adv}) = \max \left[\|f_{RGB}^n - f_{RGB}^{adv}\|_2 - \|f_{RGB}^p - f_{RGB}^{adv}\|_2 + \rho, 0 \right], \quad (12)$$

ρ is a margin parameter that controls the minimum distance between the negative pair and the positive pair. It can be used to bring negative pairs closer together and push positive pairs further apart.

$$\Delta_{meta.train} = \theta \Delta'_{meta.test} + \frac{\nabla_{\eta} L_{RGB}}{\|\nabla_{\eta} L_{RGB}\|_1}, \quad (13)$$

$$\eta = \text{clip}(\eta + \alpha \cdot \text{sign}(\Delta_{meta.train}), -\varepsilon, \varepsilon). \quad (14)$$

Here, θ represents the momentum value, $\Delta'_{meta.test}$ is derived from the previous iteration, α denotes the iteration step size, and ε is the adversarial bound. The above is the optimization process for the universal perturbation η , which results in the adversarial sample:

$$x_{adv} = x + \eta. \quad (15)$$

Using adversarial samples for adversarial training, the adversarial loss is:

$$L_{adv}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \left[\sum_{i=1}^C y_i \log(f(x + \eta)) \right], \quad (16)$$

where C is the number of classes, y_i corresponds to the true label, η is the generated adversarial perturbation.

To ensure consistency between features before and after perturbation, we define the alignment loss as follows:

$$L_{align} = \|f(x) - f(x_{adv})\|_2^2, \quad (17)$$

where $f(x)$ represents the feature vector of the original sample x , and $f(x_{adv})$ represents the feature vector of the perturbed sample x_{adv} . This loss term enforces the model to produce similar feature representations for the original and perturbed samples.

The overall loss function for the meta-training phase, denoted as $L_{meta.train}$:

$$L_{meta.train} = L_c + L_{adv} + L_{align}. \quad (18)$$

Here, L_c promotes discriminative feature learning via contrastive learning, L_{adv} encourages robust feature learning, and L_{align} ensures feature consistency between original and perturbed samples. The combination of these losses enables robust cross-modal alignment and generalization.

Meta-Testing

This step is similar to meta-training. We utilize the sketch ReID dataset to learn the perturbation η with the our loss functions:

$$L_s(f_{RGB}^{adv}) = \max \left[\|f_s^n - f_{RGB}^{adv}\|_2 - \|f_s^p - f_{RGB}^{adv}\|_2 + \rho, 0 \right], \quad (19)$$

where f_s^n, f_s^p and f_{RGB}^{adv} denote the feature vectors of the sketch image and the RGB adversarial sample in the sketch ReID dataset, respectively.

$$\Delta_{meta.test} = \theta \Delta_{meta.train} + \frac{\nabla_{\eta} L_s}{\|\nabla_{\eta} L_s\|_1}, \quad (20)$$

$$\eta = \text{clip}(\eta + \alpha \cdot \text{sign}(\Delta_{meta.test}), -\varepsilon, \varepsilon). \quad (21)$$

Here, $\Delta_{meta.train}$ derived from step of meta-training. The main difference compared to the previous step lies in the perturbation applied to the input and the gradients related to momentum.

Similarly, the adversarial loss L_{adv} in the meta-testing phase can be obtained from equations (8) and (9). The alignment loss L_{align} can then be derived from equation (10). Therefore, the loss in the meta-testing phase can be expressed as follows:

$$L_{meta.test} = L_c + L_{adv} + L_{align}. \quad (22)$$

Meta-Update

The final step is based on the losses from the previous two steps. Specifically, our ultimate loss function can be expressed as:

$$L_{total} = L_{meta.train} + L_{meta.test}. \quad (23)$$

The former aims to learn foundational knowledge through meta-training to build the model's cross-modal feature representation capability, while the latter seeks to effectively transfer this acquired knowledge to the sketch ReID task.

Experiments

Experimental Settings

We conduct experiments on three benchmark datasets. PKU-Sketch (Pang et al. 2018) contains hand-drawn sketches and corresponding RGB images of pedestrians under diverse poses and backgrounds, targeting sketch-based cross-modal ReID. Market-Sketch-1K (Lin et al. 2023) introduces artist subjectivity to simulate subtle sketch variations, offering a realistic testbed for generalization under modality and style discrepancies. Market-1501 (Zheng et al. 2015) is a large-scale RGB-only dataset containing 1,501 identities captured from six camera views. In this work, we use it as a single-modal source dataset for transfer learning. Unless otherwise specified, all experiments are conducted with Market-1501 as the default auxiliary training set to support cross-modal transfer learning.

For evaluation, we follow standard ReID protocols (Pang et al. 2018; Zhang et al. 2022b), reporting Rank-1 accuracy and mean Average Precision (mAP). Rank-1 measures the

(a) Market-Sketch-1K Dataset			(b) PKU-Sketch Dataset		
Methods	Rank-1 (%)	mAP (%)	Methods	Rank-1 (%)	mAP (%)
DSCNet (Zhang et al. 2022a)	13.8	14.7	AFLNet (Pang et al. 2018)	34.0	-
DEEN (Zhang and Wang 2023)	12.1	12.6	CCSC (Zhang et al. 2022b)	86.0	83.7
UNIReID (Chen, Ye, and Jiang 2023)	14.6	13.2	SS-reID (Lin et al. 2023)	78.0	78.7
SS-reID (Lin et al. 2023)	18.1	19.6	UNIReID (Chen, Ye, and Jiang 2023)	91.4	91.8
AIO (Li et al. 2024)	15.4	13.9	DALNet (Liu et al. 2024b)	90.0	86.2
Ours	21.3	22.7	Ours	92.7	92.3

Table 1: Comparison with SOTA methods. Left: Market-Sketch-1K (single-query); Right: PKU-Sketch. Market1501 is not used as auxiliary data to ensure fair comparison.

Dataset	Method	80% data		60% data		40% data	
		Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)
PKU-Sketch	UNIReID	83.3 ± 0.8	78.0 ± 0.3	69.6 ± 0.2	64.3 ± 0.7	44.0 ± 0.7	40.1 ± 0.5
	Ours	90.0 ± 0.2	85.0 ± 0.4	85.1 ± 0.3	80.3 ± 0.5	68.2 ± 0.3	64.1 ± 0.2
Market-Sketch-1k	SS-reID	12.2 ± 0.4	11.0 ± 0.3	9.8 ± 0.5	9.2 ± 0.4	6.5 ± 0.6	6.0 ± 0.4
	Ours	21.5 ± 0.3	23.1 ± 0.3	18.2 ± 0.4	19.7 ± 0.3	14.8 ± 0.5	16.3 ± 0.4

Table 2: Comparison of cross-modal performance under limited data. Experiments are conducted on the PKU-Sketch and Market-Sketch-1K datasets, with the latter evaluated under the single-query setting. Each episode adopts a 5-way setup by sampling 5 sketch identities and using their corresponding RGB samples as support. Meta-test identities are disjoint from meta-train to ensure fair evaluation.

retrieval accuracy of the top-1 result, while mAP reflects the overall ranking quality of the retrieved results.

Our experiments are conducted on 2 NVIDIA A40 GPUs. In our KTC module setup, we configure the hyperparameters as follows: we set the maximum iteration count $\text{max_iter} = 10$, with a margin $\rho = 0.5$ and a learning rate $\alpha = \epsilon/10$. We use stochastic gradient descent (SGD) with a momentum factor $\theta = 0.9$ to update η . For L_∞ bounded attacks, ϵ is set to 8. Given the imbalance in dataset sizes between meta-training and meta-testing, we conduct one meta-test after every 10 meta-training iterations, with a batch size $N_b = 32$.

Comparison with State-of-the-Art Methods

Tab.1 presents a comparison of our method with SOTA approaches on the Market-Sketch-1K and PKU-Sketch datasets. For fairness, no single-modal data is used in this evaluation. On Market-Sketch-1K, our method achieves 21.3% Rank-1 and 22.7% mAP, significantly outperforming SS-reID (Lin et al. 2023) (18.1% Rank-1, 19.6% mAP). This gain stems from the proposed AA and KTC modules, which improve cross-modal alignment and robustness, especially in handling subjective variations from different sketch artists. On PKU-Sketch, our method reaches 92.7% Rank-1 and 92.3% mAP, surpassing UNIReID (91.4% Rank-1, 91.8% mAP). While DALNet improves sketch-photo alignment via auxiliary modalities, our method explicitly tackles the two key factors in domain adaptation theory—domain discrepancy and perturbation stability—through alignment augmentation and adversarial perturbations. This enables our model to generalize effectively across datasets. In the Appendix, we further evaluate the model’s generalization to unseen domains via cross-dataset testing between PKU-

Sketch and Market-Sketch-1K. When trained on PKU and tested on Market-Sketch-1K, our method achieves 15.4% Rank-1 and 12.7% mAP, outperforming SS-reID (11.2% / 10.6%). In the reverse direction, our method reaches 38.1% Rank-1 and 32.6% mAP vs. SS-reID’s 32.8% / 29.2%. These results demonstrate that KTCAA exhibits stronger cross-domain generalization even without auxiliary labels, reinforcing its effectiveness under few-shot and domain-shift scenarios.

Transfer Learning Under Data Scarcity

Tab.2 reports the performance of KTCAA under varying data availability on the PKU-Sketch and Market-Sketch-1K datasets. All models are pre-trained on Market-1501 and fine-tuned using 80%, 60%, and 40% of the training data. Results are reported as the mean ± standard deviation over multiple runs. On PKU-Sketch, KTCAA consistently outperforms UNIReID across all data scales. For example, at the 40% level, KTCAA achieves substantially higher Rank-1 accuracy and mAP compared to UNIReID. On the more challenging Market-Sketch-1K, our method also shows clear advantages over SS-reID in all settings. These results demonstrate KTCAA’s superior data efficiency and cross-modal generalization under few-shot conditions.

Ablation Experiment

Tab.3 presents the ablation results on Market-Sketch-1K and PKU-Sketch. Starting from a baseline, we progressively introduce the AA and KTC modules along with their corresponding losses. On Market-Sketch-1K, the AA module injects localized modality transformations to guide progressive distribution alignment, improving Rank-1 accuracy

Baseline	+AA	+KTC	+L _c	+L _{adv}	+L _{align}	Market-sketch-1K		PKU-Sketch	
						Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)
✓						12.1	15.3	68.2	69.0
✓		✓	✓	✓		17.4	18.7	76.5	75.1
✓	✓		✓			18.0	19.9	80.9	80.2
✓	✓	✓	✓	✓		23.2	23.5	87.9	86.4
✓	✓	✓	✓	✓	✓	24.9	26.2	93.6	92.9

Table 3: Evaluation of each component of the proposed method on the Market-Sketch-1K and PKU-Sketch datasets.

Baseline	+CLIP-Align	+KTCAA	Rank-1 (%)	mAP (%)
✓			16.9	17.4
✓	✓		21.4	20.8
✓		✓	24.7	25.2
✓	✓	✓	25.1	26.5

Table 4: Comparison of cross-modal alignment performance based on the SS-reID baseline on the Market-Sketch-1K dataset.

from 12.1% to 18.0%. Introducing the KTC module (with adversarial perturbation and L_{adv} loss) further improves robustness to modality shifts, boosting Rank-1 to 23.2%. The full model, with both KTC and the alignment loss L_{align} , achieves the highest performance with 24.9% Rank-1 and 26.2% mAP, demonstrating the complementary benefits of AA and KTC for alignment and generalization. To better understand the mechanism behind these gains, we conduct additional experiments in the Appendix. First, we compare global versus local augmentation and show that the AA module achieves better generalization than a global sketch-style transformation. Next, we perform a sensitivity analysis on two key KTC hyperparameters: the adversarial bound ϵ and the triplet margin ρ . Results indicate that $\epsilon = 8$ and $\rho = 0.5$ provide the best trade-off between performance and efficiency. Finally, we report training times under different settings. While KTC increases computational cost due to adversarial optimization, it does not affect inference time, and the performance gains justify the additional training effort.

Comparison of Cross-Modal Alignment

On the Market-Sketch-1K dataset, we evaluate the performance of the SS-reID baseline model and its integration with the CLIP alignment module (CLIP-Align). For a fair comparison, the competing methods were pre-trained on the Market1501 dataset.

As shown in Tab.4, the SS-reID baseline achieves a Rank-1 accuracy of 16.9% and an mAP of 17.4%, highlighting its limitations in cross-modal alignment. Incorporating the CLIP-Align module improves the Rank-1 accuracy to 21.4% and mAP to 20.8%. With our proposed KTCAA framework, these metrics are further boosted to 25.1% Rank-1 and 26.5% mAP, achieving the highest performance. This result highlights KTCAA’s effectiveness in simultaneously reducing domain discrepancy and improving perturbation stability, thereby achieving stronger cross-modal alignment and



Figure 2: Qualitative comparison of retrieval results between SS-reID (a) and our KTCAA (b) on the Market-Sketch-1K dataset. Each row shows top-6 results for a sketch query. Green and red boxes indicate correct and incorrect matches, respectively.

generalization, even when integrated with existing alignment methods.

Visualization

To qualitatively assess the effectiveness of KTCAA, we visualize top-6 retrieval results in Fig.2. Compared to SS-reID, KTCAA retrieves more accurate matches with better consistency across pose, viewpoint, and modality variations, demonstrating stronger cross-modal alignment and discrimination. In addition, t-SNE visualizations provided in the Appendix show that KTCAA yields more compact identity clusters and better separation across modalities, further validating its alignment capability at the feature level.

Conclusion

This work introduces KTCAA, a theoretically inspired framework for few-shot sketch-based person re-identification. Inspired by generalization theory, KTCAA addresses domain discrepancy and perturbation sensitivity via two modules: AA, which injects sketch-style cues for cross-modal alignment, and KTC, which enhances robustness by reducing worst-case feature shifts. Jointly trained under a meta-learning paradigm, KTCAA enables effective RGB-to-sketch transfer and achieves state-of-the-art performance in low-data scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276222.

References

- Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning*, 79: 151–175.
- Chen, C.; Ye, M.; and Jiang, D. 2023. Towards modality-agnostic person re-identification with descriptive query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15128–15137.
- Chen, C.; Ye, M.; Qi, M.; and Du, B. 2023. Sketch-Trans: Disentangled prototype learning with transformer for sketch-photo recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Chen, Q.; Quan, Z.; Zheng, Y.; Li, Y.; Liu, Z.; and Mozerov, M. G. 2024. MSIF: multi-spectrum image fusion method for cross-modality person re-identification. *International Journal of Machine Learning and Cybernetics*, 15(2): 647–665.
- Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.
- Ge, J.; Li, H.; and Zhang, Y. 2024. Robust discriminative and modal-consistent feature learning for fine-grained sketch-based image retrieval. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, 1–8.
- Gong, Y.; Hou, Y.; Zhang, C.; and Jiang, M. 2024a. Beyond augmentation: Empowering model robustness under extreme capture environments. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.
- Gong, Y.; Huang, L.; and Chen, L. 2021. Eliminate deviation with deviation for data augmentation and a general multi-modal data learning method. *arXiv preprint arXiv:2101.08533*.
- Gong, Y.; Huang, L.; and Chen, L. 2022. Person re-identification method based on color attack and joint defence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4313–4322.
- Gong, Y.; Zhong, Z.; Qu, Y.; Luo, Z.; Ji, R.; and Jiang, M. 2024b. Cross-modality perturbation synergy attack for person re-identification. *Advances in Neural Information Processing Systems*, 37: 23352–23377.
- Gu, L.; Liu, J.; Liu, X.; Wan, W.; and Sun, J. 2024. Entropy-optimized deep weighted product quantization for image retrieval. *IEEE Transactions on Image Processing*, 33: 1162–1174.
- Gu, L.; Shen, X.; Sun, J.; Liu, Y.; Li, J.; Li, Z.; Cheung, S.-C. S.; and Wan, W. 2025. Dual Prototypes-Based Personalized Federated Adversarial Cross-Modal Hashing. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Guo, J.; Zhu, X.; Zhao, C.; Cao, D.; Lei, Z.; and Li, S. Z. 2020. Learning meta face recognition in unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6163–6172.
- Huang, W.; Liang, J.; Shi, Z.; Zhu, D.; Wan, G.; Li, H.; Du, B.; Tao, D.; and Ye, M. 2025. Learn from Downstream and Be Yourself in Multimodal Large Language Model Fine-Tuning. In *ICML*.
- Li, H.; Ye, M.; Zhang, M.; and Du, B. 2024. All in one framework for multimodal re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17459–17469.
- Li, S.; Leng, J.; Gan, J.; Mo, M.; and Gao, X. 2025a. Shape-centered representation learning for visible-infrared person re-identification. *Pattern Recognition*.
- Li, S.; Leng, J.; Kuang, C.; Tan, M.; and Gao, X. 2025b. Video-Level Language-Driven Video-Based Visible-Infrared Person Re-Identification. *IEEE Transactions on Information Forensics and Security*.
- Li, Y.; Sun, Y.; Qin, Y.; Peng, D.; Peng, X.; and Hu, P. 2025c. Robust Duality Learning for Unsupervised Visible-Infrared Person Re-Identification. *IEEE Transactions on Information Forensics and Security*, 20: 1937–1948.
- Lin, K.; Wang, Z.; Wang, Z.; Zheng, Y.; and Satoh, S. 2023. Beyond Domain Gap: Exploiting Subjectivity in Sketch-Based Person Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2078–2089.
- Liu, M.; Wang, F.; Wang, X.; Wang, Y.; and Roy-Chowdhury, A. K. 2024a. A Two-Stage Noise-Tolerant Paradigm for Label Corrupted Person Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7): 4944–4956.
- Liu, X.; Cheng, X.; Chen, H.; Yu, H.; and Zhao, G. 2024b. Differentiable Auxiliary Learning for Sketch Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3747–3755.
- Liu, Z.; Zhang, L.; and Yang, Y. 2020. Hierarchical bi-directional feature perception network for person re-identification. In *Proceedings of the 28th ACM international conference on multimedia*, 4289–4298.
- Lu, Y.; Yang, M.; Peng, D.; Hu, P.; Lin, Y.; and Peng, X. 2025. LLaVA-ReID: Selective Multi-image Questioner for Interactive Person Re-Identification. *arXiv preprint arXiv:2504.10174*.
- Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1979–1993.
- Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1765–1773.
- Pang, L.; Wang, Y.; Song, Y.-Z.; Huang, T.; and Tian, Y. 2018. Cross-domain adversarial feature learning for sketch re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, 609–617.
- Pang, Z.; Zhao, L.; Liu, Y.; Sharma, G.; and Wang, C. 2025a. Joint Augmentation and Part Learning for Unsupervised Clothing Change Person Re-Identification. *IEEE Transactions on Information Forensics and Security*.

- Pang, Z.; Zhao, L.; Liu, Y.; Wang, C.; and Sharma, G. 2025b. Robust Labeling and Invariance Modeling for Unsupervised Cross-Resolution Person Re-Identification. *IEEE Transactions on Image Processing*.
- Pu, R.; Sun, Y.; Qin, Y.; Ren, Z.; Song, X.; Zheng, H.; and Peng, D. 2025. Robust Self-Paced Hashing for Cross-Modal Retrieval with Noisy Labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19969–19977.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, 1842–1850. PMLR.
- Sun, X.; Leng, X.; Wang, Z.; Yang, Y.; Huang, Z.; and Zheng, L. 2024a. CIFAR-10-Warehouse: Broad and More Realistic Testbeds in Model Generalization Analysis. In *The Twelfth International Conference on Learning Representations*.
- Sun, X.; Yao, Y.; Wang, S.; Li, H.; and Zheng, L. 2024b. Alice Benchmarks: Connecting Real World Re-Identification with the Synthetic. In *The Twelfth International Conference on Learning Representations*.
- Sun, X.; and Zheng, L. 2019. Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 608–617.
- Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1199–1208.
- Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J. C.; Murino, V.; and Savarese, S. 2018. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31.
- Wang, J.; Gong, T.; and Yan, Y. 2024. Semi-supervised Prototype Semantic Association Learning for Robust Cross-modal Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 872–881.
- Yang, B.; Chen, J.; Ma, X.; and Ye, M. 2023. Translation, association and augmentation: Learning cross-modality re-identification from single-modality annotation. *IEEE Transactions on Image Processing*.
- Yang, B.; Chen, J.; and Ye, M. 2024. Shallow-Deep Collaborative Learning for Unsupervised Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16870–16879.
- Yang, B.; Ye, M.; Chen, J.; and Wu, Z. 2022a. Augmented Dual-Contrastive Aggregation Learning for Unsupervised Visible-Infrared Person Re-Identification. In *ACM MM*, 2843–2851.
- Yang, M.; Huang, Z.; Hu, P.; Li, T.; Lv, J.; and Peng, X. 2022b. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14308–14317.
- Yang, M.; Huang, Z.; and Peng, X. 2024. Robust object re-identification with coupled noisy labels. *International Journal of Computer Vision*, 132(7): 2511–2529.
- Yang, M.; Li, Y.; Hu, P.; Bai, J.; Lv, J.; and Peng, X. 2022c. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 1055–1069.
- Zhang, L.; Liu, Z.; Zhang, W.; and Zhang, D. 2023. Style uncertainty based self-paced meta learning for generalizable person re-identification. *IEEE Transactions on Image Processing*, 32: 2107–2119.
- Zhang, P.; Wang, Y.; Liu, Y.; Tu, Z.; and Lu, H. 2024. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *CVPR*, 17117–17126.
- Zhang, Y.; Kang, Y.; Zhao, S.; and Shen, J. 2022a. Dual-semantic consistency learning for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 18: 1554–1565.
- Zhang, Y.; and Wang, H. 2023. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2153–2162.
- Zhang, Y.; Wang, Y.; Li, H.; and Li, S. 2022b. Cross-compatible embedding and semantic consistent feature construction for sketch re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3347–3355.
- Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, 1116–1124.
- Zheng, Y.; Zhong, B.; Liang, Q.; Zhang, S.; Li, G.; Li, X.; and Ji, R. 2025. Towards universal modal tracking with on-line dense temporal token learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhong, Z.; Zheng, L.; Cao, D.; and Li, S. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1318–1327.
- Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 598–607.
- Zhu, F.; Zhu, Y.; Jiang, X.; and Ye, J. 2022. Cross-domain attention and center loss for sketch re-identification. *IEEE Transactions on Information Forensics and Security*, 17: 3421–3432.