FedSH: Towards Privacy-Preserving Text-Based Person Re-Identification

Wentao Ma, Xinyi Wu, Shan Zhao, Tongqing Zhou, Dan Guo, *Member, IEEE*, Lichuan Gu, Zhiping Cai, *Member, IEEE*, and Meng Wang, *Fellow, IEEE*

***Abstract*—Text-based person re-identification (ReID) has enabled canonical applications in searching for and tracking targets from large-scale surveillance images with textual descriptions. Yet, existing text-based person ReID systems employ centralized model training that gathers images captured by different institutes' cameras into one place, which poses severe privacy threats to sensitive institutional information. This work is then devoted to exploring privacy-preserving text-based person ReID and proposes the framework of FedSH by tailoring the federated learning paradigm for distributed searching knowledge extraction. Specifically, FedSH resolves the local model generalization and entity boundary obscuring limitations, caused by inner-institute data homogeneity and inter-institute data heterogeneity, via building multi-granularity feature representation and a semantically self-aligned network. Meanwhile, it reduces the communication burden introduced by the embedding for multiple modals by updating common representation subspaces during federated learning. Experimental results on two public benchmarks demonstrate that our method can achieve at most 16.47% and 16.02% person ReID performance improvement by the Rank-1 metric, compared with 6 State-of-The-Art (SoTA) baselines and 6 ablation studies. We believe that our work will inspire the community to investigate the potential of implementing Federated Learning in real-world image retrieval and ReID scenarios.**

***Index Terms*—Text-based Person ReID, Cross-modal Retrieval, Federated Learning, Multi-granularity Representation.**

Manuscript received 20 June 2022; revised 12 February 2023, 25 June 2023, and 16 October 2023; accepted 21 October 2023. Date of publication 6 November 2023; date of current version 21 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62072465, 62102425, 62302144, and 31771679, in part by the Science and Technology Innovation Program of Hunan Province under Grants 2023RC3027 and 2022RC3061, in part by the Postgraduate Research and Innovation Project of Hunan Province under Grant CX20210080, in part by the Major Scientific and Technological Projects in Anhui Province under Grant 201903a06020009, in part by the Natural Science Foundation of Anhui Province under Grants 2108085MF209 and 2308085MF217, in part by the Natural Science Research Project of Anhui Provincial Department of Education under Grants KJ2020A0107 and 2022AH050889, and in part by the University Synergy Innovation Program of Anhui Province under Grants GXXT-2022-046, GXXT-2022-055, and GXXT-2022-040. The Associate Editor coordinating the review of this manuscript and approving it for publication was Professor Jianguo Zhang. *(Corresponding authors: Shan Zhao; Tongqing Zhou.)*

Wentao Ma and Lichuan Gu are with the School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei 230036, China (e-mail: wtma@ahau.edu.cn; glc@ahau.edu.cn).

Xinyi Wu, Tongqing Zhou, and Zhiping Cai are with the College of Computer, National University of Defense Technology, Changsha 410073, China (e-mail: wuxinyi17@nudt.edu.cn; zhoutongqing@nudt.edu.cn; zpcai@nudt.edu.cn).

Shan Zhao, Dan Guo, and Meng Wang are with the School of Computer and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: zhaoshan@hfut.edu.cn; guodan@hfut.edu.cn; eric.mengwang@gmail.com).

Digital Object Identifier 10.1109/TMM.2023.3330091

I. Introduction

Person re-identification (ReID) has enabled the extraction of activity insights from large-scale deployments of urban cameras [1], [2], [3]. To date, such a technique finds plenty of applications in intelligent video surveillance, urban safety, and citizen services. Informally, ReID searches for images of some specific individuals from a large image gallery based on inter-correlations. In this way, we can depict the temporal-spatial trace of the target.

Although seems like a simple image retrieval task, real-world practices of person ReID are hindered by the absence of visual cues for those targets. For example, when attempting to find criminals or suspects, we usually don't have any recent and clear image records on them, even though we may have witnesses. Fortunately, verbal descriptions for the target person are generally available (e.g., from testimonies). In order to automatically establish the correlation between verbal description and widely sensed images, the literature has recently proposed the feasible paradigm of text-based person ReID [4], [5], [6], [7], [8], [9], [10], [11], [12]. Namely, identifying the target from the spatial image stream with only a text description on it. Technically, as given by the pioneering work [4], CNN-RNN networks are trained towards dedicated encoders for image -text feature representation.

A serious concern on existing text-based person ReID design is the potential breach of visual privacy with tons of images that are continuously captured in the field. For training a well-performed ReID model, the images of individuals at different physical regions and different times are gathered to a centralized server [2], [13]. Yet, in the urban scale, the distributed camera devices belong to different institutes (e.g., campus, government, and company) [3]. Opening local surveillance data will leak the sensitive information of the staff, activity, and infrastructure that is treasured by the institutes [14]. Consequently, ensuring privacy preservation in text-based person ReID is critical for both the applications and the contributions of the images.

Motivated by these issues, this work proposes to exploit the recent distributed learning paradigm, Federated Learning (FL) [2], [3], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], for privacy-preserving text-based person ReID. Intuitively, by dividing the centralized model training process into FL's iterative local training and global aggregation, we can extract identification knowledge from local images without exposing their visual contents. Actually, we note that there exists a few works built on

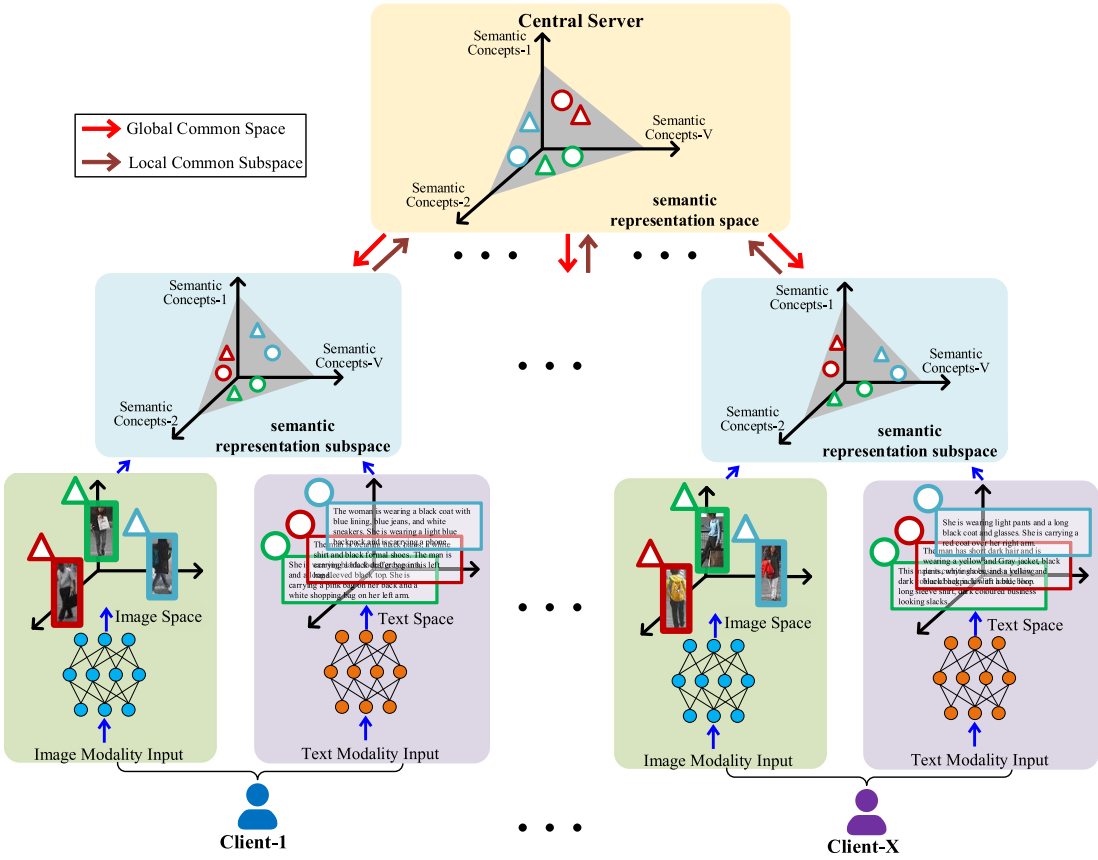

Fig. 1. Illustration of our proposed FedSH model. In the semantic common space of feature representations, the same color indicates associated semantics and the shapes represent the modalities (*i.e.*, image and text). Each round of interaction training between the client and the central server consists of the following steps: (1) The server sends a global common space to each client participating in collaborative training. (2) The clients train a model of text-based person ReID using local data to learn the local common subspace. (3) Each participating training client uploads its common subspace. (4) The server aggregates common subspace via a weighted average to obtain a new global common space.

the idea of using FL for person ReID privacy (e.g., FedReID [14] and FedUReID [20]). However, they are designed with the strong assumption that visuals on the target are available during searching, *i.e.*, image-based person ReID. In contrast, we focus on the more general scenarios of text-based person ReID in the FL setting and propose the dedicated framework of FedSH (shorted for Federated Learning-based person searching), as shown in Fig. 1. Basically, FedSH works in three steps: (1) Local training of clients for building the text-based person ReID model on local data; (2) FL global aggregation that combines the semantic common subspace of clients; (3) Local updating that replaces the local common subspace with the aggregated global common space.

Alongside the practical benefits of existing FL-based person ReID designs and privacy gain, FedSH faces three aspects of challenges:

- Feature homogeneity of each client's local images limits representation learning. Existing person ReID techniques exploit merely the global feature representation during training. Nevertheless, local datasets are constituted with images captured under the same region, which have somewhat similar, if not indistinguishable, global features. As a result, the trained local models struggle to generalize effectively to global searching tasks.
- Feature distribution heterogeneity of images on different clients harms the identification accuracy. The surveillance samples from various clients (institutes) exhibit differences in visual characteristics (e.g., uniformed personnel in the government region *v.s.* uniformed students in the campus gallery). Training with similar text descriptions independently, different clients' images on different identities may be embedded into nearby subspaces, in order to be aligned with the text. This will blur the boundary between different identities, leading to inaccurate searching.
- Cross-modal interactions incur doubled communication costs on one energy-sensitive local client. In cross-modal FL, the local model, which typically includes DNNs for feature embedding of at least two modalities (e.g., textual modality and visual modality), results in doubled transmission updates between clients and the server. Such costs on communication would significantly add up energy burden of the client (e.g., IoT cameras).

To cope with those tricky challenges, FedSH first introduces a part-level feature alignment that parses images (using smaller

kernels) and text descriptions (using attention mechanism) at fine granularity and combines it with global-level feature extraction during representation learning. In this way, FedSH tunes the representation to more different details at local training for enhanced generalization (Challenge 1). It then jointly considers inter- and intra-sample distance to align samples for both visual similarity and visual-textual correlation, addressing the heterogeneity of feature distribution across local clients' datasets (Challenge 2). Finally, we analyze the redundancy in raw model updating and propose to exchange only the semantic common subspace to improve efficiency (Challenge 3). Our contributions can be summarised as follows:

- To the best of our knowledge, our study is the *first* attempt to protect the privacy of text-based person ReID via FL deployment. A FedSH framework is presented for the distributed cross-modal learning process and key challenges are analyzed.
- FedSH accommodates feature representation performance and efficiency with a semantically self-aligned network, multi-granularity feature representation, and semantic common subspace extraction.
- On two public benchmarks of text-based person ReID, we evaluate FedSH by comparison with 6 SoTA baselines and a series of ablation studies. The results show at most 16.47% and 16.02% person ReID performance improvement.

The rest of this paper is organized as follows. First, we briefly review the related work in Section II and Section III introduces text-based person ReID under the FL setting. Then, we present the experimental settings and results in Section IV and V. Finally, conclusions are given in Section VI.

II. Related Work

In this section, we will briefly review the most relevant study to our work, including federated learning, image-based person ReID, and text-based person ReID.

A. Federated Learning

FL, firstly proposed by Google in [15], is an emerging distributed training technique, which does not require centralized data training compared to traditional machine learning. It is able to coordinate decentralized clients to learn a shared global model while guaranteeing each participant's original data secure [2], [3]. Since the early success of FL in image classification [16], [22], it has been widely accepted in other fields. For example, researchers in personalized search [19], recommendation system [17], multi-modal representation learning [21], [23] and information retrieval [14], [20], [21], [24], [25], [26] has used FL to complete downstream tasks.

Caldas et al. [16] present LEAF, an open-source benchmark framework for FL datasets, which focuses on computer vision and natural language processing tasks, and achieves promising performance. To address the problems of sensitive information leakage on personalized search tasks, Yao et al. [19] propose a personalized search framework with privacy-preserving, called FedPS, which is the first attempt to tackle the dilemma of privacy protection in personalized search tasks. Guo et al. [17] propose an edge-accelerated perception FL framework for user point-of-interest recommendation. This model can not only improve model performance but also protect data security. To handle the ubiquitous non-iid problem in data generated by most mobile scenarios, Liu et al. [18] propose a distributed-aware FL to alleviate the heterogeneity of data distribution.

Moreover, FL has also been applied in person ReID and cross-modal retrieval. For privacy protection of multimodal data, Xiong et al. [23] design a multimodal FL framework, called MMFed. The MMFed uses the co-attention mechanism to capture the complementary information between different modalities of each local client, which can improve the performance and protect the privacy of multimodal data. Inspired by the success of FL, Zong et al. [21] present a new framework, named FedCMR, to provide a solution for cross-modal retrieval in distributed data. The FedCMR model completes one FL mechanism for cross-modal retrieval through steps such as aggregating the common subspace, smooth transition from the global common space to the local common subspace, and two-stage reinforcement training. Zhuang et al. [14] propose FedReID, which is the first effort to implement FL in person ReID. It can realize efficient retrieval of personal images and protect privacy of personal. Also, considering that FedReID heavily relies on data labels during client training, in another work, Zhuang et al. [20] present FedUID, which realizes efficient image-based person ReID system without labels while protecting privacy. Different from these works, inspired by [21], [23], we focus on text-based person ReID tasks under the FL paradigm with a clear definition and a heuristic solution to this task.

B. Image-Based Person ReID

Image-based person ReID has enabled the extraction of actionable insights from large-scale deployments of urban cameras [1], [2], [3], [13], [27]. To date, such a technique has been applied to intelligent video surveillance [14], [28] and content-based video retrieval [13], [29], [30], [31]. Informally, ReID searches for images of some specific individuals from a large image gallery based on the inter-correlation. Some of the existing person ReID models borrow architectures for image recognition tasks to extract discriminative features for persons [13], [29] and the others use effective attention mechanisms to suppress irrelevant features (e.g., image background and noise) to enhance feature discrimination [32], [33], [34]. Several works [35], [36], [37] fuse semantic and detail representations for person ReID tasks. Chen et al. [35] propose a novel salience-guided cascaded suppression network for person ReID, termed SCSN, which enables the model to mine and fuse diverse salient features to improve the discriminability of feature representation. Leveraging the advantage of both CNNs and Transformers, Zhang et al. [37] propose a hierarchical aggregation transformer framework to better integrate fine-grained semantic information for image-based person ReID.

Nevertheless, since these images contain personal information and identification, existing training models of image-based

person ReID require the centralization of data, which raises the potential risk of compromising personal information. Thus, some recent works [14], [20], [24], [25], [26] have applied FL to image-based person ReID, to protect personal privacy. Among them, Zhuang et al. [14] propose FedReID, which is the first attempt, to realize efficient retrieval and protect the privacy of personally sensitive information. Also, considering that the FedReID heavily relies on data labels during local client training, in another work, Zhuang et al. [20] present FedUID, which can realize an efficient image-based person ReID system without labels while protecting privacy. Sun et al. [24] survey a selective knowledge aggregation framework (SKG) that focuses on a decentralized image-based person ReID. SKG improves personalization by local batch processing for each domain, focusing on the normalization layer, and improving model generalization by using local normalization mechanisms. To handle the poor generalization capability, Yang et al. [25] propose a Domain and Feature Hallucinating (DFH) for person ReID, called FedDG, which enables the local model to see as diverse samples as possible. In order to handle the heterogeneity in FL person ReID, Wu et al. [26] decentralize learning of non-shared private training data distributed on multiple user sites in independent multi-domain label space. Although these methods enable clients to achieve image-based person ReID with superior performance and privacy-preserving, the problem setting has major limitations in practice. This is because that the real-world practice of person ReID is hampered by the absence of visual cues for those targets, making the target image. For example, when attempting to find criminals or suspects, we may not have any photo records on them.

C. Text-Based Person ReID

The technique of image-based person ReID has plenty of applications in real-world scenarios. However, in some real-world practices, the target query image is not always available. As a result, to adapt the diverse person ReID application scenarios, several investigations [4], [5], [6], [7], [8], [9], [10], [11], [12], [38], [39], [40], [41], [42], [43] focus on text-based person ReID. These work which can be broadly divided into global-level alignment matching, local-level alignment matching, and multi-granularity alignment matching according to the image-text cross-modal matching manner.

Early text-based person ReID models utilize the CNN-RNN network to learn global-level cross-modal feature representation and make efforts to design a more appropriate objective loss function, like [4], [5], [38]. In particular, Li et al. [4] is the first to propose a task of text-based person ReID, while collecting a benchmark dataset, called CUHK-PEDES. Yet, the above methods all leverage global-level feature representation for text-based person ReID, which can hardly capture those unique local semantic representations. To tackle this issue, the local-level alignment matching methods have been explored [6], [7]. Wherein, Aggarwal et al. [6] propose to narrow the modality gap between image and text by introducing person attribute representations as the complementation of global-level features. In addition, more and more work has explored multi-granularity alignment matching [8], [9], [11], [12], [40], [41], [42], [43], [44]. Jing et al. [40] propose a multi-granularity alignment model based on the attention mechanism, which achieves semantic overlay from coarse-to-fine via a fine-grained alignment component and a coarse-fine alignment component. Niu et al. [39] present a multi-level alignment method including global-global alignment, global-local alignment, and local-local pairwise alignment, then fusing their results. A simple yet effective framework for text-based person ReID, called TIPCB, is explored in [12], which realizes efficient retrieval by means of multi-level alignment matching. Farooq et al. [11] design a novel multi-layer network, called AXM-Net, which aligns cross-modal feature representations by suppressing the background information to highlight the semantic information induced by the personal image. To enable better cross-modal alignment, Ding et al. [8] propose a semantically self-aligned network model, which bridges the gap between modalities by fusing global-level matching and part-level matching.

Although existing text-based person ReID models have achieved remarkable progress, yet, there is still a gap between the task setting and real-world scenarios. For example, training the models of text-based person ReID, which requires centralizing a large amount of data, may raise potential privacy risks. Thus, this work aims to implement text-based person ReID under the FL with privacy preserved.

III. Methodology

This section introduces the proposed FedSH, a novel framework for implementing FL to text-based person ReID, which includes Problem Definition (Section III-A), Local Client Training (Section III-B) and Central Server Aggregation (Section III-C).

A. Problem Definition

Suppose there are X local clients, indicated as $C = \{C^1, C^2, \ldots, C^X\}$, and all clients have local data denoted as $D = \{D^1, D^2, \ldots, D^X\}$, including the personal images and corresponding text descriptions, where D^x is the set of samples for x-th client. FedSH aims to learn a model of text-based person ReID that can correctly search the corresponding personal image by a text description, in each local client. All participating clients need to collaborate to train the model without centralizing the data. For readability and clarity, some of the notations adopted in our paper and their definitions are described in Table I.

B. Local Client Training

Text-based person ReID is a special application of cross-modal retrievals, like existing image-text matching [9], [11], [42], which is devoted to learning a semantic representation. In such a space, where the items of different modalities can be compared to each other. Nevertheless, the inherent modality gap between images and texts makes this task challenging. To bridge this gap, inspired by recent multi-granularity alignment methods [8], [41], [45], we leverage a semantically multi-granularity self-aligned network for text-based person ReID in

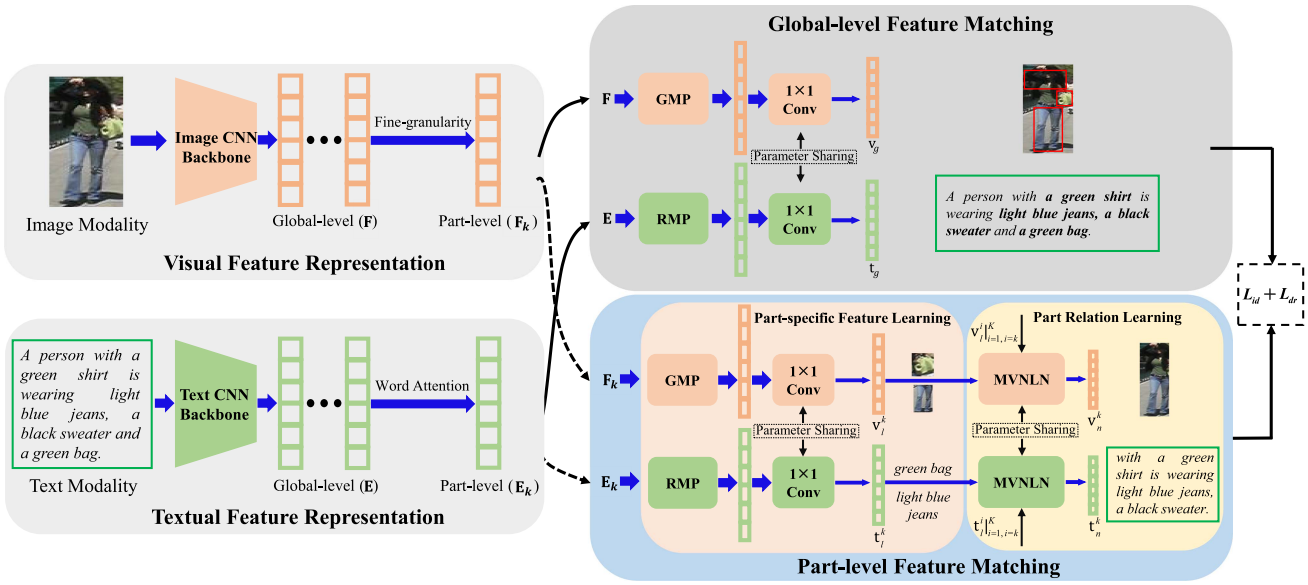

Fig. 2. Architecture of text-based person ReID in local clients. The model is built on ResNet-50 and Bi-LSTM backbones, which extract global-level and part-level features from image modality and text modality, respectively. For the Global-level Feature Matching, we adopt a weight-sharing strategy on the last 1×1 Conv layer, which aligns the feature representations from the two modalities more tightly in terms of semantics. The Part-level Feature Matching is composed of a Part-specific Feature Learning (PFL) module and a Part Relation Learning (PRL) module. Wherein, the PFL module and the PRL module enable the model to automatically extract part-level features from both modalities and capture relationships between different parts to establish more precise semantic correspondences with noun phrases, respectively.

TABLE I
KEY NOTATIONS

Notations	Description
X	X represents the number of local client.
C^x	C^x indicates the x-th client.
D^x	D^x indicates the number of data for x-th client.
T	One text description T that consists of n words $\{c_1, c_2, \cdots, c_n\}$.
$\mathbf{F}$, $\mathbf{E}$	$\mathbf{F}$ and $\mathbf{E}$ represent the feature map of image and the word embedding of text description, respectively.
(I^+, T^+)	(I^+, T^+) indicates the positive image-text pairs in a batch size.
(I^+, T^-), (T^+, I^-)	(I^+, T^-) and (T^+, I^-) indicate the negative image-text pairs in a batch size.
S_g, S_l, and S_n	S_g, S_l, and S_n represent global-level feature representation similarity, part-level feature representation similarity, and part-level feature representation similarity yielded by MVNLN of one image-text pair, respectively.
R	R Indicates the number of communication rounds between the central server and the clients.
b	b indicates the number of samples in a batch size.

each client. As shown in Fig. 2, the framework of the local client model mainly contains three modules: (1) Visual-Textual Feature Representation, (2) Global-level Feature Matching, and (3) Part-level Feature Matching.

1) Visual-Textual Feature Representation: The prevailing pre-trained ResNet and Bi-LSTM are leveraged as the backbone to extract visual and textual feature representations, respectively.

Visual Feature Representation: We extract the personal image feature map, denoted as $\mathbf{F}$, to learn the global-level feature representation of images. For the part-level ones, as semantic concepts

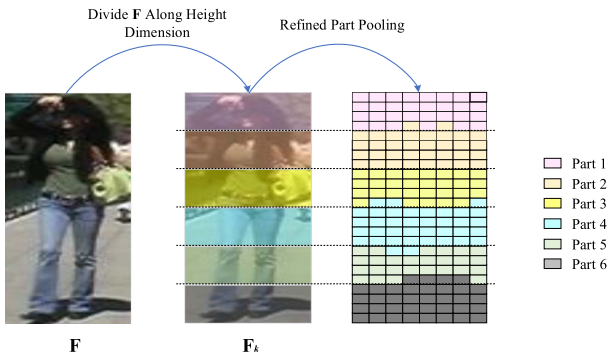

Fig. 3. Illustration of the generation process for fine granularity visual features.

in the images are complex, it commonly exists in benchmark datasets [4], [8] that a noun or phrase is able to describe local attributes of multiple similar identities. Namely, semantic boundaries between texts to describe different identities are potentially ambiguous. Hence, inspired by [8], [46], we mine fine granularity visual features, as shown in Fig. 3. Specifically, we design a simple and uniform partitioning strategy without external cues for part-level visual feature extraction. *First*, considering the model size, algorithm complexity, and part feature's quality, we divide the feature map $\mathbf{F}$ evenly into K non-overlapping parts in the height dimension, to get $\mathbf{F}_k$ $(1 \leq k \leq K)$. However, there exist many outliers, while designated to a specified horizontal stripe (part) during training, which is more similar to another part. To tackle this issue, we further adopt the Refined Part Pooling [46] for correcting inconsistency of within-part, which is aimed at assigning all the column vectors according to their similarities to each part. In this way, the outliers will be

relocated and we can get the part features with adaptive similarity. Following the experience of previous work [8], we set K=6 by default, to realize the trade-off between discriminant ability of features and optimization of model performance.

Textual Feature Representation: Text descriptions contain all the information about the appearance of a personal image. Specifically, the whole sentence describes the global-level feature representation in images, while part-level ones contain multiple entities and actions. Therefore, multi-granularity representation learned from global-to-part is beneficial to accurately and comprehensively understand the semantic information of text descriptions. Given a text description T that consists of n words $\{c_1, c_2, \ldots, c_n\}$, to capture the salient features between words in a sentence, we utilize the Bi-LSTM to generate a sequence $\mathbf{E} = [e_1, e_2, \ldots, e_n]$ of rich semantic contextual-aware word embeddings:

$$\overrightarrow{w}_i = \overrightarrow{\text{LSTM}}\left(\mathbf{W_c} c_i, \overrightarrow{w}_{i-1}; \overrightarrow{\theta}\right) \tag{1}$$

$$\overleftarrow{w}_i = \overleftarrow{\text{LSTM}}\left(\mathbf{W_c} c_i, \overleftarrow{w}_{i+1}; \overleftarrow{\theta}\right) \tag{2}$$

$$e_i = \left(\overrightarrow{w}_i + \overleftarrow{w}_i\right)/2 \tag{3}$$

where $\mathbf{W_c}$ is the word embedding matrix, $\overrightarrow{\theta}$ and $\overleftarrow{\theta}$ are parameters in the two LSTMs.

2) Global-Level Feature Matching: Our method is to learn a semantic common space of visual-textual representation in which items of different modalities can be compared with each other. To obtain the global-level feature representation of image-text, inspired by [8], [44], we conduct Global Max Pooling (GMP) and Row-wise Max Pooling (RMP) on image feature map $\mathbf{F}$ and text description word embedding $\mathbf{E}$, respectively. Then we project the obtained feature representations into a semantic common space.

$$\mathbf{v}_g = \mathbf{W}_g GMP(\mathbf{F}) \tag{4}$$

$$\mathbf{t}_g = \mathbf{W}_g RMP(\mathbf{E}) \tag{5}$$

where $\mathbf{W}_g$ is a shared 1×1 Conv layer, $\mathbf{v}_g$ and $\mathbf{t}_g$ represent global-level feature representations of image and text, respectively. Finally, the similarity score between global-level feature representations is denoted as:

$$S_g = \frac{\mathbf{v}_g^\top \mathbf{t}_g}{\|\mathbf{v}_g\| \|\mathbf{t}_g\|} \tag{6}$$

3) Part-Level Feature Matching: In multi-granularity alignment for text-based person ReID [7], [8], [9], [39], [40], [41], [42], [43], part-level feature representation has been widely-adopted. As a result, inspired by prior work [8], we design a Part-specific Feature Learning (PFL) module and a Part Relation Learning (PRL) module as part branches to learn the part-level feature representation.

Part-specific Feature Learning: To ensure the consistency of part-level feature representations, Word Attention Module (WAM) is adopted to infer the semantic correspondence between word-embedding and parts. Meanwhile, we predict the probability that the i-th word belongs to the k-th part as follows:

$$s_i^k = \sigma\left(\mathbf{W}_p^k e_i\right) \tag{7}$$

where $\mathbf{W}_p^k$ is a Conv layer and the σ represents the sigmoid function. Hence, the text description for the k-th part is denoted as:

$$\mathbf{E}_k = \left[s_1^k e_1, s_2^k e_2, \ldots, s_n^k e_n\right] \tag{8}$$

Then, we obtain the visual feature representations for the k-th part by feeding $\mathbf{F}_k$ into one GMP layer and one 1×1 Conv layer. Similar to generating global-level feature representations, we conduct RMP on $\mathbf{E}_k$ to generate the k-th part-level textual feature representations and feed it to the same 1 × 1 Conv layer as $\mathbf{F}_k$. Formally,

$$\mathbf{v}_l^k = \mathbf{W}_l^k GMP\left(\mathbf{F}_k\right) \tag{9}$$

$$\mathbf{t}_l^k = \mathbf{W}_l^k RMP\left(\mathbf{E}_k\right) \tag{10}$$

where $\mathbf{W}_l^k$ is a shared 1×1 Conv layer to project the k-th visual feature representation $\mathbf{v}_l^k$ and the k-th part textual feature representation $\mathbf{t}_l^k$ into a semantic common space. Thus, the similarity score of one image-text pair at part-level feature representation can be formulated as:

$$S_l = \frac{\mathbf{v}_l^\top \mathbf{t}_l}{\|\mathbf{v}_l\| \|\mathbf{t}_l\|} \tag{11}$$

Part Relation Learning: Although several works [8], [46], [47] have demonstrated that the equal partition strategy on image feature map $\mathbf{F}$ is effective. Yet, since the semantic concepts are complex, it commonly exists in benchmarks that one phrase may cover multiple semantically similar parts. Thus, to explore the correlation between evenly divided parts, following prior work [8], we leverage a Multi-view Non-local Network (MVNLN) on the k-th part feature representation. The similarity between $\mathbf{v}_l^k$ and $\mathbf{v}_l^i (i \neq k)$ in a semantic common space based on multi-view projections can be denoted as:

$$S_{ki} = \frac{\beta_k\left(\mathbf{v}_l^k\right)^\top \phi_i\left(\mathbf{v}_l^i\right)}{\left\|\beta_k\left(\mathbf{v}_l^k\right)\right\| \left\|\phi_i\left(\mathbf{v}_l^i\right)\right\|} \tag{12}$$

where $\beta_k(\mathbf{v}_l^k) = \mathbf{W}_\beta^k \mathbf{v}_l^k$, $\phi_i(\mathbf{v}_l^i) = \mathbf{W}_\phi^i \mathbf{v}_l^i$, and $\mathbf{W}_\beta^k$ and $\mathbf{W}_\phi^i$ are both a Conv layer.

Then, the interactive relationship between the feature representation of k-th part and K-1 part can be denoted as:

$$\alpha_{ki} = \frac{\exp\left(S_{ki}\right)}{\sum_{i=1, i\neq k}^{K} \exp\left(S_{ki}\right)} \tag{13}$$

the calculated α_{ki} is then used to aggregate the K-1 part features:

$$\mathbf{v}_{l_{in}}^k = \mathbf{W}_\gamma^k \left(\sum_{i=1, i\neq k}^{K} \alpha_{ki} \phi_i\left(\mathbf{v}_l^i\right)\right) \tag{14}$$

Accordingly, the part-level visual feature representation yielded by the MVNLN module can be formulated as:

$$\mathbf{v}_n^k = \mathbf{W}_n^k \left(\mathbf{v}_l^k + \mathbf{v}_{l_{in}}^k\right) \tag{15}$$

where $\mathbf{W}_\gamma^k$ and $\mathbf{W}_n^k$ are both a Conv layer.

For capturing the correlations of part-level textual feature representations, similar to the visual modality, we reuse MVNLN on text descriptions. It is worth noting that the parameters of MVNLN are shared between the two modalities data (according to [8], [48]). Furthermore, similar to S_g and S_l, we also utilize the cosine distance to measure the similarity of one image-text pair at the feature representation yielded by MVNLN:

$$S_n = \frac{\mathbf{v}_n^\top \mathbf{t}_n}{\|\mathbf{v}_n\| \|\mathbf{t}_n\|} \tag{16}$$

where $\mathbf{v}_n$ and $\mathbf{t}_n$ represent the K part-level feature representations of image and text generated by MVNLN, respectively.

Finally, the overall similarity can be described as:

$$S = (S_g + S_l + S_n)/3 \tag{17}$$

4) Objective Function: For cross-modal retrieval, the positive representations of image-text pairs are pulled close while the negative ones are pushed apart by pair-wise ranking loss [8], [9], [11], [42], [45]. However, direct deployment of pair-wise ranking loss is unreliable for the image-text retrieval model [8], [48], because the ambiguous semantic boundaries between texts to describe different images can bring spurious correlation between texts and their non-corresponding images. We design a dual-ranking loss (L_{dr}) based on basic pair-wise ranking loss (L_1) and weak supervision terms (L_2).

For L_1, we basically follow the pair-wise ranking loss formulation as in several prior work [8], [45], [48]. Given a quadric input (I^+, T^+, I^-, T^-), for each positive pair (I^+, T^+), we find the hardest negatives within a batch size (I^+, T^-) and (T^+, I^-), and push them away from the positive pair beyond a pre-defined margin Δ. L_1 can be written as,

$$\begin{aligned} L_1 = &\max\left[\Delta - S\left(I^+, T^+\right) + S\left(I^+, T^-\right)\right] \\ &+ \max\left[\Delta - S\left(T^+, I^+\right) + S\left(T^+, I^-\right)\right] \end{aligned} \tag{18}$$

where $S(\cdot,\cdot)$ represents the distance measurement criterion (cosine similarity is used here). Given a query image I^+, the similarity score of the semantic matching text should be higher. When giving a query text T^+, we expect the relevant image to be ranked higher.

For L_2, the weakly supervised terms consist of a text description of one image and another image with the same identity, L_2 can be written as,

$$\begin{aligned} L_2 = &\mu \cdot \max\left[\Delta_1 - S\left(I^+, T^{+'}\right) + S\left(I^+, T^-\right)\right] \\ &+ \mu \cdot \max\left[\Delta_1 - S\left(T^{+'}, I^+\right) + S\left(T^{+'}, I^-\right)\right] \end{aligned} \tag{19}$$

where $T^{+'}$ is a text description of an image with the same identity as I^+, Δ_1 represents a pre-defined margin, and μ is the weight of weak regulation. Hence, L_{dr} can be formulated as:

$$L_{dr} = L_1 + L_2 \tag{20}$$

Meanwhile, the objective loss function (L_{dr} and L_{id} [48]) is utilized to optimize the Global-level Feature Matching and Part-level Feature Matching, respectively. During the inference test, we only use the S_g as the final similarity of an image-text pair, which can also reduce the amount of calculation.

C. Aggregation for Federated Learning

In this subsection, we describe the major step of the proposed federated aggregation, FedSH, including the central server aggregation framework and the aggregation mechanism.

1) Aggregation Framework: Standard FL aggregation [14], [15], [20] requires bi-directional communication between multiple clients and the server to realize cooperative training. However, when the structure of the local client model is complex with a large number of parameters, aggregation would usually lead to large communication costs during the client-server cooperative training process. In particular, the proposed cross-modal person ReID model would unfortunately incur doubled communication costs on energy-sensitive local clients, considering that local models are with two-tower structure for feature embedding of both the image and text modality. Such costs on communication would significantly add up energy burden on the clients.

To reduce the training communication costs, inspired by [21], we leverage the semantic common space to share the knowledge among all clients. Specifically, the central server collects semantic common subspaces and combines them through (21), searching for a globally consistent potential common space W_r.

$$W_r = \sum_{x=1}^{X} \frac{q_r^x}{d} W_r^x \tag{21}$$

where X is the number of clients being involved in cooperative training during the r-th round of communication between the server and clients, W_r^x is the updated semantic common subspace of local client x, and $d = \sum_{x=1}^{X} q_r^x$ is the total amount of data on all clients. The weight of the x-th client depends on the relative ratio between the number of data in the current client q_r^x and d. Algorithm 1 presents the process of FedSH.

2) Aggregation Mechanism: This work employs two mainstream FL aggregation methods, FedAvg [49] and FedProx [50].

FedAvg: We aggregate all local models on the server with a weighted average. The weight of the x-th client depends on the relative ratio between the number of datasets in the current client q_r^x and d. There are X clients participating in collaborative training and the FL aggregation equation is shown in (21).

FedProx: It is an optimization framework that is essentially a generalization and re-parameterization of *FedAvg*, which is believe to be able to tackle the systems and statistical heterogeneity inherent in FL.

IV. Experimental Settings

In this section, we will briefly introduce the public benchmark datasets, evaluation metrics, baselines, and implementation details.

A. Datasets

CUHK – PEDES [4]. It consists of 40206 pedestrian images and 80412 text descriptions for 13003 persons with two captions per image. Following the official evaluation guideline, the training set includes 11003 identities with 34054 images and 68108 text descriptions, while the validation set and testing set both include data for 1000 identities with 3078 and 3074

Algorithm 1: : *The proposed FedSH.*

Input: Local client training epoch E, batch size B, number of communication round R, number of selected clients X, and learning rate ξ. All clients have private data $D = \{D^1, D^2, \ldots, D^X\}$.

Output: The semantic common subspace W_r^x of client x and the global semantic common space W_r in the r-th communication round.

Sever Aggregation:
  Initialize the global semantic common space W_0;
  for each communication round r from 0 to R **do**
    $C_r \leftarrow$ randomly select X clients during the r-th round communication;
    for each client $x \in C_r$ **in parallel do**
    $W_{r+1}^x \leftarrow$ **Client** (W_r);
  $W_{r+1} \leftarrow \sum_{x \in C_t} \frac{D^x}{|D^{C_r}|} W_{r+1}^x$;
  return W_R.
Client (W_r):
  if r==0 then
    Initialize local client model $W_r^x \leftarrow W_r$ with local data D^x;
  $B \leftarrow$ (split data D^x into batches of size B);
  for each local client training epoch $0 \sim E$ **do**
    for $b \in B$ **do**
      $W_{r+1}^x \leftarrow W_r^x - \xi \nabla \mathcal{L}(W_r^x, b)$;
  return W_{r+1}^x.

pedestrian images, respectively. The image and text descriptions of the testing set constitute the gallery set and the query (probe) set, respectively. All experiments are implemented based on this train-test split.

ICFG − PEDES [8]. It contains 4102 identities with 54522 pedestrian images and each image has one text description. Following the general evaluation guideline, we divide ICFG-PEDES into a training set of 34674 image-text pairs on 3102 identities and a testing set of 19848 image-text pairs on 1000 identities.

B. Evaluation Metrics

We evaluate the proposal in terms of the ReID performance, communication costs, and computation costs. For ReID performance, we use the public cumulative-match-characteristic (CMC) curve, namely, Recall at H (Rank-H), for evaluation. Given a text description as the query, Rank-H ranks images in the semantic common space by estimating similarity. Rank-H is the possibility that the true match appears in the top-H of the rank list, where we utilize H = 1, 5, 10 (following the settings of prior ReID works). In addition, for a more comprehensive assessment, we also leverage the sum of all Rank-H to measure the overall retrieval quality:

$$\mathbf{sum} = (Rank - 1) + (Rank - 5) + (Rank - 10) \quad (22)$$

For communication costs, we follow the basic facts in FL, namely, more communication rounds lead to higher communication costs under a fixed-size training model. In FedSH, we only upload the common subspace, rather than the entire raw client model.

For computation costs, we measure the computation cost by the number of local epochs. Although the computation cost of similarity measurement of semantic space varies within rounds, it is negligible compared with the training of backbone. Therefore, similar to [14], [20], [21], [22], we approximate the computation costs by the number of epochs.

C. Baselines

We propose to evaluate the proposal's performance in terms of *local training* (to test the effectiveness of feature representation design) and *federated learning* (to test the effectiveness on the optimization and aggregation design), respectively.

- Baselines for local training performance: *Raw* utilizes the image-text global-level feature matching component and L_{id} component for local training. We further construct a series of variants by combining *Raw* with the part-level feature matching (*PFL* or *PRL*) to get *Raw+PFL* and *Raw+PRL*, and adopting L_1 loss and L_2 loss to get *Raw+*L_{dr}. Details on these components are introduced in Section III-B. In addition to the above ablation setup, we also introduce 6 popular models for performance evaluation, which are Dual-Path [48], MPA+CMPC [38], MIA [39], SCAN [51], ViTAA [7], and SSAN [8].
- Baselines for FL performance: We test FedSH's performance under different updating policies (i.e., whole parameter uploading *v.s.* semantic common subspace uploading) and different aggregation algorithms (i.e., FedAvg *v.s.* FedProx).

D. Implementation Details

The experiments are conducted on Windows with Intel (R) Core (TM) i9-10940X CPU@3.30 GHz, 128 GB RAM, and two Nvidia GeForce RTX-2080Ti GPUs. We implement FedSH in Python using EasyFL based on the PyTorch framework. The backbone structures of image modality and text modality model are the popular pre-training ResNet and Bi-LSTM, respectively. During training, we employ ADAM [52] as our optimizer with learning rate $= 1 \times 10^{-3}$, batch size $= 32$, local epoch $= 40$, and total training round $= 30$.

V. Experimental Results and Analysis

We conduct extensive experiments for text-based person ReID tasks on two public benchmarks, aiming at testing the performance of the local client training design (RQs #1∼#3 in Section V-A) and the FL aggregation design (RQs #4∼#6 in Section V-B):

- **RQ1:** Is the overall performance of our local client model superior to the SoTA baselines?
- **RQ2:** How do different components (e.g., the *PFL*, the *PRL*, and L_2) affect our local client model?

TABLE II
SERIES OF ABLATION STUDIES EACH COMPONENT, AND THE TEXT-BASED PERSON REID COMPARISON WITH SOTA MODELS ON CUHK-PEDES AND ICFG-PEDES DATASETS

Raw	Components/Methods			CUHK-PEDES				ICFG-PEDES			
	PFL	*PRL*	L_{dr}	Rank-1	Rank-5	Rank-10	**sum**	Rank-1	Rank-5	Rank-10	**sum**
✓				18.50	39.17	50.88	108.55	24.50	46.16	57.90	128.56
✓	✓			31.39	55.19	66.86	153.44	21.85	39.79	49.65	111.29
✓	✓	✓		35.28	59.73	70.25	165.26	25.15	44.54	55.68	125.37
✓			✓	50.46	74.41	83.37	208.24	51.70	72.90	80.78	205.38
✓	✓		✓	55.45	78.34	86.77	220.56	52.40	71.36	78.47	202.23
✓	✓	✓	✓	**60.87**	**80.82**	**87.61**	**229.30**	**55.01**	**72.75**	**79.48**	**207.24**
Dual-Path [51]				44.40	66.26	75.07	185.73	38.99	59.44	68.41	166.84
CMPA+CMPC [39]				49.37	-	79.27	-	43.51	65.44	74.26	183.21
MIA [40]				53.10	75.00	82.90	211.00	46.49	67.14	75.18	188.81
SCAN [54]				55.86	75.97	83.69	215.52	50.05	69.65	77.21	196.91
ViTAA [7]				55.97	75.84	83.52	215.33	50.98	68.79	75.78	195.55
SSAN [8]				59.36	79.06	85.56	223.98	52.63	72.05	78.90	203.58

Abbreviations "*PFL*", "*PRL*" and "L_{dr}" represent part-level feature learning module, part-level relation learning module, and loss function, respectively. Here, "-" denotes that no experimental results with the same settings are available.

- **RQ3:** What is the visualization of image-text feature representation?
- **RQ4:** How does the performance of FedSH vary under different federated aggregations?
- **RQ5:** What is the impact of different subspace sizes on FedSH's performance?
- **RQ6:** What is the convergence speed of FedSH?

A. ReID Performance of Local Client Training

To demonstrate the performance of our local client model, we conduct a series of ablation studies and compare them with SoTA baseline models, the detailed results are shown in Table II.

1) Basic Performance Comparisons (RQ1 & RQ2): We compare the performance of the client model gained with local client training with 6 SoTA baselines and several ablation designs. Results are shown in Table II.

For baseline comparison, on CUHK-PEDES, the client model consistently achieves the best performance compared to its counterparts on all indicators. To be specific, its performance is further enhanced to 229.30 in sum, which indicates a significant improvement (+43.57), compared with Dual-Path. On ICFG-PEDES, for each metric, our client model outperforms the best competitors, SCAN [51] and ViTAA [7], and the overall retrieval quality reflected by the sum is also boosted by a large margin (relative +10.33 and +11.69, respectively). Especially, compared with the competitive competitor, SSAN [8], our client model can achieve the most outstanding results on all indicators.

Furthermore, a series of ablation studies are conducted to investigate the contributions of different components in our client model, we compare the model with its six counterparts, which are *Raw* and five variations of our local client model: *Raw*+*PFL*, *Raw*+*PFL*+*PRL*, *Raw*+L_{dr}, *Raw*+*PFL*+L_{dr} and *Raw*+*PFL*+*PRL*+L_{dr}. The experimental results are shown in Table II, from which we can gain the following observations:

- The use of *Raw*+*PFL*, *Raw*+*PFL*+*PRL* and *Raw*+L_{dr} improves *Raw*'s performance on the sum indicator by 44.89, 56.71, and 99.69 points on CUHK-PEDES. In contrast, on ICFG-PEDES, we can see that the use of *Raw*+*PFL* degrades *Raw* when it is equipped with *PFL*, as the number of embedded components increases, while our client model exceeds the *Raw* in the sum. The above comparisons demonstrate that *PFL*, *PRL* and L_2 are effective for learning the cross-modal image-text representation, resulting in better alignment.
- Meanwhile, we equip *Raw* with *PFL* and L_{dr}, namely, *Raw*+*PFL*+L_{dr}, which respectively boost the Rank-1 performance by 36.95 on CUHK-PEDES, while on ICFG-PEDES by 27.90. Note that, for SoTA baselines, Dual-Path and CMPA+CMPC, only equipping L_{dr} can achieve competitive performance while integrating *Raw* with *PFL*, *PRL* or L_{dr} can consistently facilitate better performance than the SoTA ones with a preferable margin in all metrics.
- Finally, by equipping *Raw* with all components, our client model achieves 60.87 and 55.01 of Rank-1, and 229.30 and 207.24 of sum on both datasets, respectively.

As discussed above, we conclude that the major gain comes from the multi-granularity feature matching mechanism, which enhances the complementarity of semantic representations at global-to-part, thus narrowing the modality gap between image and text.

2) Visualization of Image-Text Feature Representations (RQ3): To visually investigate the effectiveness of our local client model, we adopt the t-SNE to embed the feature representations of images-text into a 2-D plane.

As shown in Figs. 4(a) and 5(a), we can see that the inter- and intra-class samples from both image and text are hard to be distinguished in the semantic common space of the *Raw* output. Fig. 4(b)∼(c) and Fig. 5(b)∼(c) demonstrate that, as *Raw* is equipped with *PFL* and *PRL*, the model is enforced to compact the relevant image-text pairs and scatter irrelevant ones. From Figs. 4(d) and 5(d), we can see that by equipping the *Raw* with all components, the model can effectively categorize the feature representations of different modalities into several semantically distinct clusters. Moreover, the distribution of image and text is well-mixed and difficult to categorize. This means that our client model can greatly lessen the modality gap between image and text.

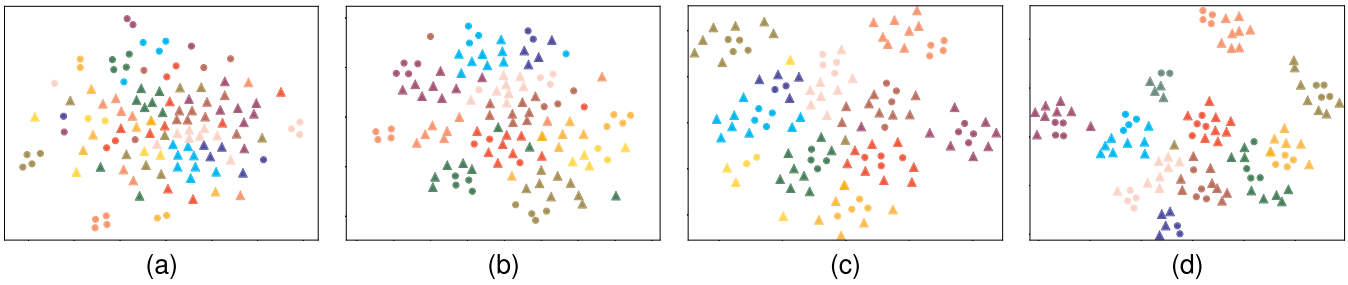

Fig. 4. Visualization semantic common space for the ten category data (including image modality and text modality) from CUHK-PEDES by using the t-SNE [53]. The same color indicates relevant semantics, the shapes represent different modalities. (a), (b), (c), and (d) are feature representations of image-text extracted via *Raw*, *Raw+PFL*, *Raw+PFL+PRL*, and *Raw+PFL+PRL+*L_{dr}, respectively.

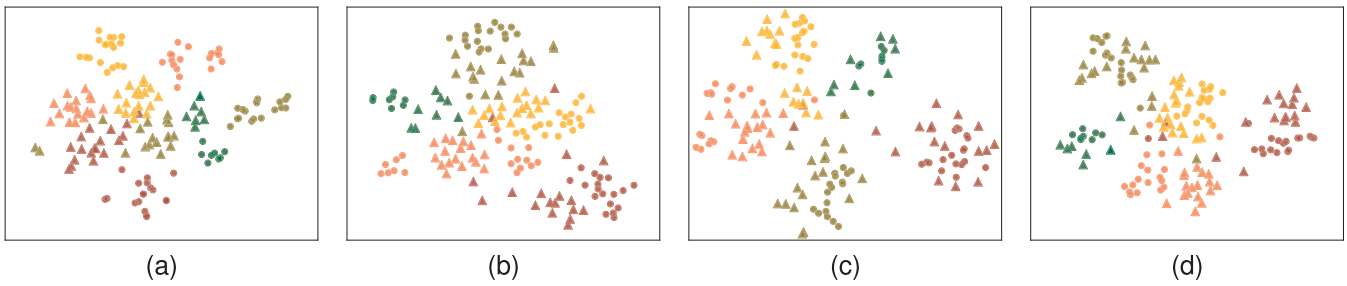

Fig. 5. Visualization semantic common space for the five category data (including image modality and text modality) from ICFG-PEDES by using the t-SNE [53]. The same color indicates relevant semantics, the shapes represent different modalities. (a), (b), (c), and (d) are feature representations of image-text extracted via *Raw*, *Raw+PFL*, *Raw+PFL+PRL*, and *Raw+PFL+PRL+*L_{dr}, respectively.

B. ReID Performance Under FL Aggregation

This work is the first attempt to protect privacy in text-based person ReID under FL deployment. A FedSH framework is proposed for the distributed cross-modal learning process and key challenges are analyzed. Therefore, we conduct extensive experiments on two public benchmark datasets to demonstrate the feasibility and application prospects of our method by comparing local client training with FedAvg and FedProx.

1) Comparisons on Federated Aggregation (RQ4): To demonstrate the effectiveness and feasibility of the proposal, we respectively implement FedSH under FedAvg and FedProx aggregation with the same setting, and add centralized local client training, Dual-Path and SCAN, to the comparisons. The experimental results are summarized in Table III, from which we can attain the following four observations:

- On CUHK-PEDES and ICFG-PEDES, the centralized training model of our client model has a superior performance compared with FedSH* (with FedAvg) and FedSH* (with FedProx). In particular, on CUHK-PEDES, the performance of the client model is 229.30 in sum, showing a significant improvement (+66.75 and +70.15) over FedSH* (with FedAvg) and FedSH* (with FedProx), respectively. While on ICFG-PEDES, for each metric, the centralized training model of our client model outperforms FedSH* (with FedAvg) and FedSH* (with FedProx), and the overall retrieval quality reflected by the sum shows a large margin (relative +70.7 and +30.72, respectively).

TABLE III
PERFORMANCE COMPARISON OF CENTRALIZED LOCAL CLIENT TRAINING AND FEDSH WITH BOTH DIFFERENT FEDERATED AGGREGATION APPROACHES (E.G., FEDAVG AND FEDPROX) ON CUHK-PEDES AND ICFG-PEDES

Setting	Rank-1	Rank-5	Rank-10	**sum**
CUHK-PEDES [4]				
Local Training (Dual-Path [48])	44.40	66.26	75.07	185.73
Local Training (SCAN [51])	55.86	75.97	83.69	215.52
Local Training (Ours)	60.87	80.82	87.61	229.30
FedSH* (with FedAvg)	38.02	64.23	74.15	176.40
FedSH* (with FedProx)	32.88	57.51	68.76	159.15
FedSH (with FedAvg)	54.95	76.24	84.52	215.71
FedSH (with FedProx)	**57.48**	**78.07**	**86.00**	**221.55**
ICFG-PEDES [8]				
Local Training (Dual-Path [48])	38.99	59.44	68.41	166.84
Local Training (SCAN [51])	50.05	69.65	77.21	196.91
Local Training (Ours)	55.01	72.75	79.48	207.24
FedSH* (with FedAvg)	28.49	49.25	58.80	136.54
FedSH* (with FedProx)	41.33	63.38	71.81	176.52
FedSH (with FedAvg)	50.25	69.60	77.02	196.87
FedSH (with FedProx)	**51.07**	**70.46**	**77.57**	**199.10**

Here, "FedSH*" denotes aggregating all the parameters of the entire client model (following traditional FL), while "FedSH" represents aggregating the semantic commonsubspace of each client model.

- *For two federated aggregation:* FedSH* (with FedProx) approximates FedSH* (with FedAvg) in all metrics on CUHK-PEDES, while FedSH* (with FedProx) outperforms FedSH* (with FedAvg) in each indicator on ICFG-PEDES. In contrast, FedSH (with FedProx) exceeds FedSH (with FedAvg) in all metrics when aggregating the semantic common subspace on two datasets. Since FedProx is an optimization framework, it is believed to be able to tackle

TABLE IV
ABLATION STUDY OF DIFFERENT SUBSPACE SIZES

Methods	Subspace Size			CUHK-PEDES				ICFG-PEDES			
	Backbone		Components	Rank-1	Rank-5	Rank-10	**sum**	Rank-1	Rank-5	Rank-10	**sum**
	All	Half									
FedAvg	✓		✓	38.02	64.23	74.15	176.40	28.49	49.25	58.80	136.54
	✓			45.00	70.33	80.66	195.99	36.74	59.78	68.98	165.50
		✓		54.61	76.96	84.59	216.16	50.16	69.84	77.21	197.21
		✓	✓	54.95	76.24	84.52	215.71	50.25	69.60	77.02	196.87
FedProx	✓		✓	32.88	57.51	68.76	159.15	41.33	63.38	71.81	176.52
	✓			49.92	73.49	82.46	205.87	38.51	60.34	69.80	168.65
		✓		55.42	77.23	85.35	218.00	50.68	69.87	77.43	197.98
		✓	✓	**57.48**	**78.07**	**86.00**	**221.55**	**51.07**	**70.46**	**77.57**	**199.10**

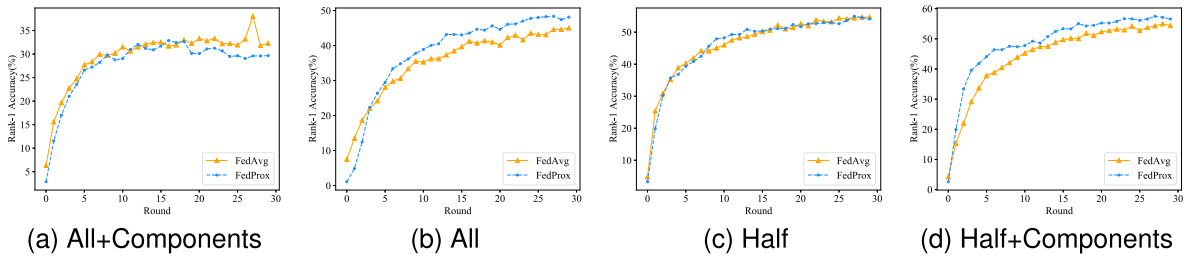

Fig. 6. Performance and convergence comparison of FedAvg and FedProx with different semantic common spaces on CUHK-PEDES, measured by Rank-1 accuracy. "All", " Half", and "Components" respectively represent the parameters of the pre-trained backbone, half of the parameters of the pre-trained backbone, and the parameters of the components.

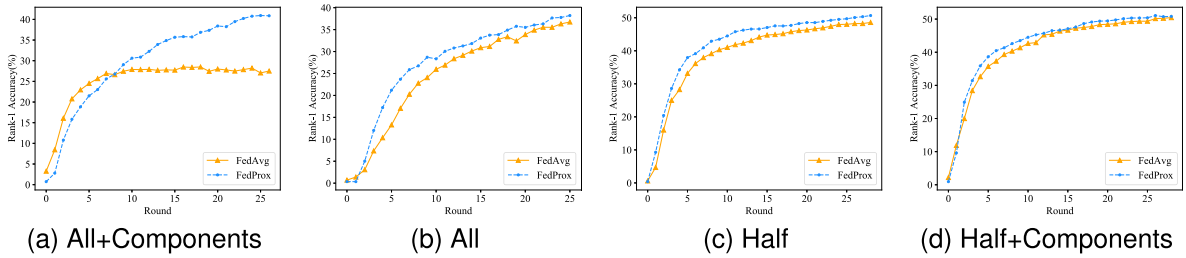

Fig. 7. Performance and convergence comparison of FedAvg and FedProx with different semantic representation spaces on ICFG-PEDES, measured by Rank-1 accuracy. "All", " Half", and "Components" respectively represent the parameters of the pre-trained backbone, half of the parameters of the pre-trained backbone, and the parameters of the components.

the systems and statistical heterogeneity inherent in federated models compared to FedAvg.

- *For the aggregation parameters:* Aggregating the entire client model is unreliable because a huge number of parameters are redundant or may contain noise. These issues will affect the performance of the federated model and also introduce computational costs. Therefore, FedSH only aggregates the semantic representation space. As shown in Table III, on two datasets, the two aggregations' results demonstrate that the FedSH paradigm is superior to the FedSH* paradigm in terms of all indicators.
- Although FedSH is inferior to the centralized training model of our client model in both federated aggregation approaches. FedSH with FedProx outperforms some SoTA local training models (Dual-Path [48] and SCAN [51]) on two datasets. In particular, FedSH (with FedProx) exceeds 6.18 and 2.19 points in sum on the CUHK-PEDES and ICFG-PEDES datasets, respectively, compared to SCAN.

2) Ablation Study of Representation Space (RQ5): To verify the influence of subspace size on FedSH performance, we conduct a series of ablation studies on the semantic common subspace. Since we leverage the size of semantic common space as a parameter carrier to control the number of parameters, "All", "Half", and "Components" respectively represent the parameters of the pre-trained two-tower backbone (*i.e.*, the parameters of ResNet-50 and Bi-LSTM), half of the parameters of the pre-trained backbone (*i.e.*, the parameters of only Bi-LSTM), and the parameters of the components we employ in our work. Experimental results on both datasets are summarized in Table IV , from which we can draw the following three observations:

- For the proposed FedSH, the model achieves the worst performance via aggregating the "All+Components". This phenomenon is consistent with our viewpoint that aggregating the entire client model is unreliable because a large number of parameters are redundant and may

contain noise, which affects the performance of the federated model and incurs communication costs. By contrast, when only "Half" is aggregated, the model leads to a moderate result, which yields 32.88 and 41.33 in Rank-1 on CUHK-PEDES and ICFG-PEDES, respectively. Obviously, the model achieves the best performance in all metrics when "Half+Components" is aggregated, obtaining 221.55 and 199.10 in the sum on two datasets.

- Meanwhile, we note that FedSH's accuracy increases steadily in each indicator, with the reduction ("All+Components"→"All"→"Half") of the number of aggregation parameters. In fact, aggregating the entire client model is unreliable for traditional FL because it is redundant or may contain noise, which makes the performance lower than expected.
- Furthermore, on two datasets, "Half+Components" is superior to "All+Component" in FedProx aggregation mechanism, and the overall retrieval quality reflected by the sum shows a large margin (relative +62.4 and +22.58, respectively). We argue that this is reasonable, because halving the parameters of the backbone significantly reduces redundancy and noise, thereby enhancing performance and lowering communication costs.

3) Model Convergence (RQ6): Convergence is a vital index of FL. We investigate the convergence to illustrate the superiority of our FedSH. As shown in Figs. 6 and 7, the FedSH is trained on CUHK-PEDES and ICFG-PEDES to test Rank-1 accuracy in 30 rounds of communication, with fixing batch size = 32.

As can be seen from Figs. 6(a) and 7(a), when "Half+Components" is aggregated, FedAvg can achieve convergence after 5 communication rounds on two datasets, while FedProx can get convergence after 5 communication rounds on CUHK-PEDES and 25 communication rounds on ICFG-PEDES. Aggregating a large number of parameters is redundant and can make the model fall into local minima, thus affecting convergence. From Figs. 6(a)–(c) and 7(a)–(c), one can see that with the reduction ("All+Components"→"All"→"Half") of aggregation parameters, the convergence speed and accuracy are steadily increasing. In particular, "Half+Components" is aggregated via FedAvg and FedProx, which can achieve convergence after 15 communication rounds and yield superior performance on two datasets.

VI. Conclusion

In this paper, we propose a novel dedicated framework for text-based person ReID in the FL setting, called FedSH. On the one hand, the FedSH enriches the semantic representation of image-text pairs by employing multi-granularity cross-modal learning from global-to-part features. On the other hand, FedSH addresses the modality gap by considering both inter- and intra-distance of samples, aligning them effectively. Finally, we propose an efficient approach where only the semantic common subspace is exchanged, enhancing the effectiveness and efficiency of federation aggregation. Extensive experiments on two public benchmarks demonstrate that FedSH is a competitive candidate for secure retrieval.

Furthermore, our model belongs to the two-tower structure, namely, employing distinct image and text encoders to match image-text pairs in the embedding semantic common space. Although this proposal achieves promising performance, yet, they only achieve sub-optimal results due to the lack of closer image-text interactions. Hence, in future work, we will explore the integration of advanced two-tower vision-language pre-training (e.g., CLIP [54] and ALIGN [55]) into image-text cross-modal retrieval tasks.

References

[1] M. Ye et al., "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[2] J. Liu et al., "From distributed machine learning to federated learning: A survey," *Knowl. Inf. Syst.*, vol. 64, pp. 885–917, 2022.

[3] D. H. Mahlool and M. H. Abed, "A comprehensive survey on federated learning: Concept and applications," in *Proc. Mobile Comput. Sustain. Inform.*, 2022, pp. 539–553.

[4] S. Li et al., "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5187–5196.

[5] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1908–1917.

[6] S. Aggarwal, V. B. Radhakrishnan, and A. Chakraborty, "Text-based person search via attribute-aided matching," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2606–2614.

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

[8] 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*.

[9] S. Zhang et al., "Text-based person search in full images via semantic-driven proposal generation," in *Proc. 4th Int. Workshop Human Centric Multimedia Anal.*, 2023, pp. 5–14.

[10] S. Zhao, C. Gao, Y. Shao, W.-S. Zheng, and N. Sang, "Weakly supervised text-based person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 11375–11384.

[11] 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.

[12] 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, 2022.

[13] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*.

[14] W. Zhuang et al., "Performance optimization of federated person re-identification via benchmark analysis," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 955–963.

[15] J. Konečný et al., "Federated learning: Strategies for improving communication efficiency," 2016, *arXiv:1610.05492*.

[16] S. Caldas et al., "Leaf: A benchmark for federated settings," 2018, *arXiv:1812.01097*.

[17] Y. Guo et al., "PREFER: Point-of-interest recommendation with efficiency and privacy-preservation via federated edge learning," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 1, pp. 1–25, 2021.

[18] B. Liu et al., "DISTFL: Distribution-aware federated learning for mobile scenarios," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 4, pp. 1–26, 2021.

[19] J. Yao, Z. Dou, and J.-R. Wen, "FedPS: A privacy protection enhanced personalized search framework," in *Proc. Web Conf.*, 2021, pp. 3757–3766.

[20] W. Zhuang, Y. Wen, and S. Zhang, "Joint optimization in edge-cloud continuum for federated unsupervised person re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 433–441.

[21] L. Zong et al., "FedCMR: Federated cross-modal retrieval," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1672–1676.

[22] H. Zeng, T. Zhou, Y. Guo, Z. Cai, and F. Liu, "FedCav: Contribution-aware model aggregation on distributed heterogeneous data in federated learning," in *Proc. Int. Conf. Parallel Process.*, 2021, pp. 1–10.

[23] B. Xiong, X. Yang, F. Qi, and C. Xu, "A unified framework for multi-modal federated learning," *Neurocomputing*, vol. 480, pp. 110–118, 2022.
[24] S. Sun, G. Wu, and S. Gong, "Decentralised person re-identification with selective knowledge aggregation," 2021, *arXiv:2110.11384*.
[25] F. Yang, Z. Zhong, Z. Luo, S. Li, and N. Sebe, "Federated and generalized person re-identification through domain and feature hallucinating," 2022, *arXiv:2203.02689*.
[26] G. Wu and S. Gong, "Decentralised learning from independent multi-domain labels for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, 2021, pp. 2898–2906.
[27] Q. Xie, W. Zhou, G.-J. Qi, Q. Tian, and H. Li, "Progressive unsupervised person re-identification by tracklet association with spatio-temporal regularization," *IEEE Trans. Multimedia*, vol. 23, pp. 597–610, 2021.
[28] H. Galiyawala, M. S. Raval, and D. Savaliya, "DSA-PR: Discrete soft biometric attribute-based person retrieval in surveillance videos," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2021, pp. 1–7.
[29] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3376–3385.
[30] N. Spolaor et al., "A systematic review on content-based video retrieval," *Eng. Appl. Artif. Intell.*, vol. 90, 2020, Art. no. 103557.
[31] H. Luo, W. Jiang, X. Fan, and C. Zhang, "STNReID: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification," *IEEE Trans. Multimedia*, vol. 22, pp. 2905–2913, 2020.
[32] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 371–381.
[33] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch dropblock network for person re-identification and beyond," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3690–3700.
[34] F. Yang, Z. Zhong, Z. Luo, S. Lian, and S. Li, "Leveraging virtual and real person for unsupervised person re-identification," *IEEE Trans. Multimedia*, vol. 22, pp. 2444–2453, 2020.
[35] X. Chen et al., "Salience-guided cascaded suppression network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3297–3307.
[36] Z. Liu, L. Zhang, and Y. Yang, "Hierarchical bi-directional feature perception network for person re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 4289–4298.
[37] G. Zhang, P. Zhang, J. Qi, and H. Lu, "Hat: Hierarchical aggregation transformers for person re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 516–525.
[38] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 686–701.
[39] 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.
[40] Y. Jing et al., "Pose-guided multi-granularity attention network for text-based person search," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11189–11196.
[41] C. Gao et al., "Contextual non-local alignment over full-scale representation for text-based person search," 2021, *arXiv:2101.03036*.
[42] C. Wang, Z. Luo, Y. Lin, and S. Li, "Text-based person search via multi-granularity embedding learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 1068–1074.
[43] A. Zhu et al., "DSSL: Deep surroundings-person separation learning for text-based person retrieval," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 209–217.
[44] W. Ma, T. Zhou, J. Qin, Q. Zhou, and Z. Cai, "Joint-attention feature fusion network and dual-adaptive NMS for object detection," *Knowl.-Based Syst.*, vol. 241, 2022, Art. no. 108213.
[45] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10635–10644.
[46] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.
[47] H. Yao et al., "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.
[48] Z. Zheng et al., "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 16, no. 2, pp. 1–23, 2020.
[49] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.
[50] T. Li et al., "Federated optimization in heterogeneous networks," *Proc. Mach. Learn. Syst.*, vol. 2, pp. 429–450, 2020.
[51] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 201–216.
[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[53] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[54] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[55] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.