Leveraging Prior Knowledge of Diffusion Model for Person Search

Giyeol Kim[1*] Sooyoung Yang[2*] Jihyong Oh[1] Myungjoo Kang[2,3] Chanho Eom[1†]

[1]GSAIM, Chung-Ang University [2]IPAI, Seoul National University
[3]Department of Mathematical Sciences and RIMS, Seoul National University

{giyeolkim, jihyongoh, cheom}@cau.ac.kr, {jimmy1016, mkang}@snu.ac.kr

https://perceptualai-lab.github.io/DiffPS/

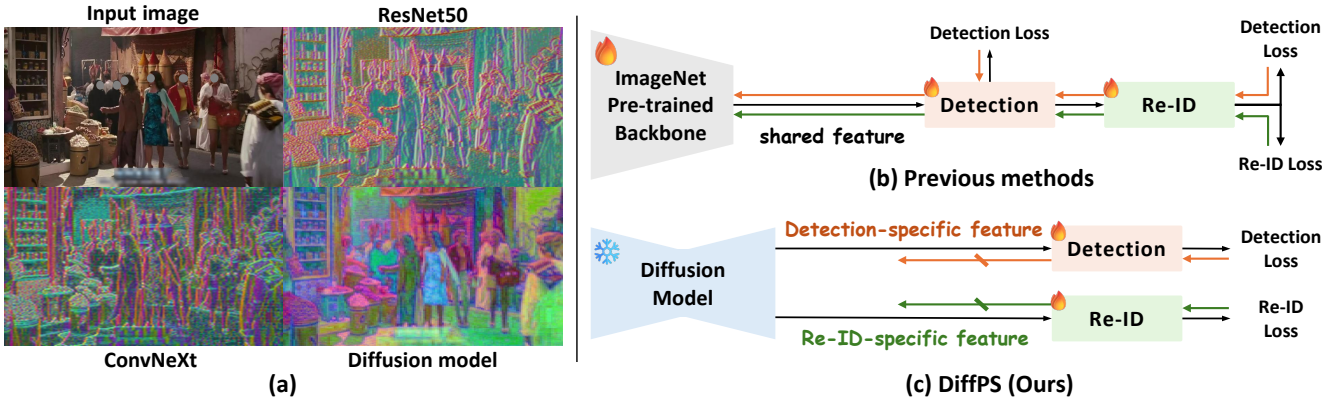

Figure 1. (a) PCA [42] visualization of feature maps, comparing widely-adopted ImageNet [9] pre-trained backbones (ResNet50 [19] and ConvNeXt [39]) in person search with a diffusion model. The diffusion features exhibit richer spatial context and fine-grained details. (b) Existing methods [3, 5, 17, 23, 26, 33, 69, 72, 73] share backbone features for detection and re-ID, causing optimization conflicts due to competing gradients (orange and green arrows). (c) Our DiffPS fully leverages diffusion priors, allowing us to freeze the backbone and utilize task-specific features for each sub-task, thereby preventing gradient interference and completely resolving the conflict.

Abstract

Person search aims to jointly perform person detection and re-identification by localizing and identifying a query person within a gallery of uncropped scene images. Existing methods predominantly utilize ImageNet pre-trained backbones, which may be suboptimal for capturing the complex spatial context and fine-grained identity cues necessary for person search. Moreover, they rely on a shared backbone feature for both person detection and re-identification, leading to suboptimal features due to conflicting optimization objectives. In this paper, we propose DiffPS (Diffusion Prior Knowledge for Person Search), a novel framework that leverages a pre-trained diffusion model while eliminating the optimization conflict between two sub-tasks. We analyze key properties of diffusion priors and propose three specialized modules: (i) Diffusion-Guided Region Proposal Network (DGRPN) for enhanced person localization, (ii) Multi-Scale Frequency Refinement Network (MSFRN) to mitigate shape bias, and (iii) Semantic-Adaptive Feature Aggregation Network (SFAN) to leverage text-aligned diffusion features. DiffPS sets a new state-of-the-art on CUHK-SYSU and PRW.

1. Introduction

Person search aims to locate and identify a query person within a gallery of uncropped scene images, consisting of two main tasks: (i) *person detection* [2, 37, 48], which localizes all person bounding boxes in each scene, and (ii) *person re-identification (re-ID)* [21, 35, 64, 78], which matches the detected person crops to the query. Early person search methods [4, 11, 16] tackle these tasks sequentially, using separate networks for detection and re-ID. While these methods achieve promising performance, their cascaded design leads to computational inefficiencies and prevents end-to-end inference. To address these limitations, recent methods [3, 5, 17, 23, 26, 33, 69, 72, 73] integrate person detection and re-ID within a unified framework, leveraging shared backbone features for both subtasks. This integration improves computational efficiency

*Equal contribution.
†Corresponding author.

and enables end-to-end inference, making it the dominant paradigm in recent research.

Despite these advantages, recent methods still face two key challenges. The first challenge is the need for a backbone with strong generalization capabilities and rich prior knowledge to effectively support both person detection and re-ID. Existing methods [3, 5, 17, 23, 26, 33, 69, 72, 73] predominantly rely on ImageNet [9] pre-trained backbones (*e.g.*, ResNet50 [19] or ConvNeXt [39]). However, as ImageNet mainly contains images with a single dominant object and simple backgrounds, these backbones are trained for category-level recognition, focusing on object presence [20, 63]. Therefore, they struggle to capture precise localization cues and instance-level discriminative features, both essential for person search in complex scenes with multiple overlapping individuals (Fig. 1 (a)). This limitation highlights the need for a pre-trained backbone capable of capturing rich contextual information and fine-grained details in challenging environments.

Another fundamental challenge is the optimization dilemma arising from the conflicting objectives between person detection and re-ID. In recent methods [3, 5, 17, 23, 26, 33, 69, 72, 73], both sub-tasks share a common backbone, which is simultaneously optimized to serve two fundamentally different goals. Person detection aims to extract features that capture general human characteristics for distinguishing individuals from the background, while person re-ID requires highly discriminative features to differentiate between specific identities. This inherent conflict forces the shared backbone to learn contradictory representations, potentially degrading the performance of both subtasks [3, 5, 17, 33, 73]. As shown in Fig. 1 (b), the backbone struggles to balance the two objectives, which hinders convergence and results in suboptimal performance due to incompatible updates within a single parameter space.

Recently, diffusion models [47, 51, 52, 54] have shown remarkable progress in image generation tasks. Trained on large-scale datasets (*e.g.*, LAION-5B [57]), they effectively capture both low-level visual features and high-level semantic relationships, enabling a comprehensive understanding of not only *what* an object is but also *where* it is located [41] (Fig.1(a)). Leveraging these capabilities, pre-trained diffusion models have been successfully applied to a variety of fundamental vision tasks, *e.g.*, image classification [6, 31], segmentation [34, 46, 49, 62, 67, 70], and semantic correspondence [40, 43, 74]. These successes highlight the potential of diffusion models as powerful general-purpose backbones, capable of handling diverse vision tasks even without fine-tuning. Inspired by these capabilities, the prior knowledge from a pre-trained diffusion model may appear well-suited for person search, due to its ability to capture both comprehensive spatial context and fine-grained details, which are essential for person search.

In this paper, we propose DiffPS (Diffusion Prior Knowledge for Person Search), a novel framework that fully exploits the prior knowledge of pre-trained diffusion models for person search (Fig. 1 (c)). To effectively harness the rich priors embedded in diffusion models, we first analyze their characteristics through four key properties: text condition, timesteps, hierarchical structure and shape bias. Based on these insights, we propose three specialized modules designed to maximize the diffusion priors for person search. First, we introduce a Diffusion-Guided Region Proposal Network (DGRPN), which refines person localization using cross-attention maps. Second, to mitigate the shape bias inherent in diffusion models, we propose a Multi-Scale Frequency Refinement Network (MSFRN) that enhances high-frequency details for improved identity discrimination. Finally, we design a Semantic-Adaptive Feature Aggregation Network (SFAN), which exploits the strong alignment between diffusion features and text embeddings to generate semantically enriched person representations. By fully leveraging diffusion priors, DiffPS eliminates the need for fine-tuning the backbone while maintaining strong performance. This allows us to freeze the backbone and extract task-specific features for person detection and re-ID, thereby completely resolving the inherent optimization conflict present in existing methods [3, 5, 26, 33, 69, 73]. As a result, our framework enables independent optimization of both tasks, preventing mutual interference and ensuring that each branch is optimized for its respective objective. To validate the effectiveness of our approach, we conduct extensive qualitative and quantitative experiments, demonstrating that DiffPS not only outperforms existing methods but also establishes diffusion models as powerful backbones for person search. The main contributions of this work are:

- We propose DiffPS, which effectively utilizes diffusion priors through three specialized modules, completely resolving the conflict between detection and re-ID.
- To the best of our knowledge, we are the first to leverage the prior knowledge of pre-trained diffusion models for person search.
- DiffPS achieves state-of-the-art performance on CUHK-SYSU [69] and PRW [77].

2. Related Work

2.1. Person Search

Person search aims to localize and identify individuals within uncropped scene images. Early methods [4, 11, 16] employ a two-stage approach using separate networks for detection and re-ID, while recent methods [3, 5, 17, 23, 26, 33, 69, 72, 73] integrate both tasks into a unified framework. However, these methods face limitations with ImageNet [9] pre-trained backbones, which struggle with contextual understanding and fine-grained feature extraction in

complex scenes. To tackle this, SOLIDER [8] introduces a pre-trained model that learns a general human representation through self-supervised learning on LUPerson [14] and PretrainPS [63] adopts a hybrid learning paradigm across multiple datasets [1, 32, 58, 66]. However, these models are primarily trained on cropped person images, which may limit their abilities to precisely locate individuals and adapt effectively across diverse and real-world scenes. In contrast, our approach leverages diffusion models pre-trained on diverse large-scale datasets, providing comprehensive spatial context and fine-grained details without domain constraints.

Another fundamental challenge lies in the optimization conflict between detection and re-ID objectives. Previous methods [3, 5, 17, 23, 26, 33, 69, 72, 73] rely on a shared backbone for both sub-tasks, resulting in competing gradients that hinder optimization and degrade performance. In contrast, DiffPS leverages the rich prior knowledge of a pre-trained diffusion model, enabling us to freeze the backbone and utilize task-specific features, thereby avoiding the need for shared backbone optimization. Although DiffPS may appear conceptually similar to DMRNet [17] in its decoupled approach, it fundamentally differs by fully resolving the conflict through diffusion priors, enabling independent task optimization without any gradient interference.

2.2. Diffusion Models for Downstream Tasks

Recently, diffusion models [22, 47, 51, 52, 54, 56], trained on large-scale datasets (*e.g.*, LAION-5B [57]), have demonstrated remarkable capabilities beyond image generation, extending to various vision understanding tasks. Several methods have leveraged diffusion features for segmentation [34, 41, 49, 62, 67, 70, 74] and classification [6, 31]. In addition, other methods [40, 43] have focused on effective feature selection within diffusion models. These methods typically utilize diffusion features through simple concatenation or aggregation. In contrast, we maximize diffusion priors via our Multi-Scale Frequency Refinement Network, which enhances high-frequency details across multi-scale features. In person retrieval-related tasks, PSDiff [25] and DenoiseRep [71] incorporate diffusion and denoising processes, but significantly differ from our approach. Specifically, PSDiff [25] adapts denoising algorithms for bounding box regression similar to DiffusionDet [7], and DenoiseRep [71] applies denoising techniques in representation learning. These methods do not leverage the internal representations of pre-trained diffusion models trained on large-scale datasets. In contrast, our method fully exploits the prior knowledge embedded in a pre-trained diffusion model, effectively adapting it for person search.

3. Method

In this section, we first introduce the diffusion model and its UNet [55] architecture (Sec. 3.1), followed by an analysis of diffusion priors for person search (Sec. 3.2). Finally, we present our proposed framework (Sec. 3.3).

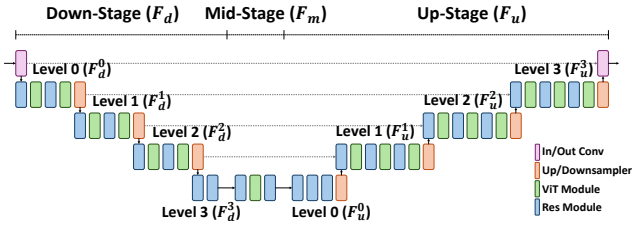

Figure 2. Detailed architecture of the UNet [55] in diffusion models [52, 54], comprising down-stage, mid-stage, and up-stage for hierarchical feature processing. (Best viewed in color.)

3.1. Preliminaries

Diffusion models. Diffusion models [22, 52, 54, 59] generate images by iteratively refining Gaussian noise through a forward and reverse process. In the forward process, noise is gradually added to a clean image $\mathbf{x}_0$ over T timesteps following a noise schedule $\{\alpha_t\}_{t=1}^{T}$, transforming it into a nearly pure noise sample $\mathbf{x}_T$ as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$. The reverse process removes noise using a denoising network $\mathcal{F}_\theta$, estimating $\epsilon \approx \mathcal{F}_\theta(\mathbf{x}_t, t)$ to reconstruct the image. For text-conditioned generation, CLIP [50] text encoder $\mathcal{T}(\cdot)$ converts a prompt $\mathbf{p}$ into an embedding $\mathbf{T}_p = \mathcal{T}(\mathbf{p}) \in \mathbb{R}^{77\times d}$, guiding the denoising process via cross-attention.

UNet Architecture in Diffusion Models. The UNet [55] architecture in diffusion models [52, 54] follows a hierarchical structure comprising three main stages, as shown in Fig. 2. The down-stage progressively reduces spatial resolution while increasing channel dimensions, whereas the up-stage restores resolution using skip connections from corresponding down-stage levels. Both the down-stage and up-stage consist of four resolution levels, each containing multiple sequential processing blocks for feature refinement. Between these stages, the mid-stage operates on the most compressed features at the bottleneck. We denote the features extracted at level l of the down-stage as $\mathbf{F}_d^l$ and those from the corresponding up-stage level as $\mathbf{F}_u^l$. Further details are provided in the Suppl. A.

3.2. Diffusion Priors for Person Search

Text condition. Pre-trained diffusion models [52, 54, 56] exhibit a strong alignment between image features and text embeddings via cross-attention [49, 70, 75, 76]. As shown in Fig. 3 (a), attention maps for different textual tokens highlight relevant image regions. This alignment provides a strong semantic prior for person localization and enables the extraction of discriminative features for body parts and clothing. By leveraging these features, the model effectively suppresses background noise and mitigates occlusions, ensuring a more robust representation for person search.

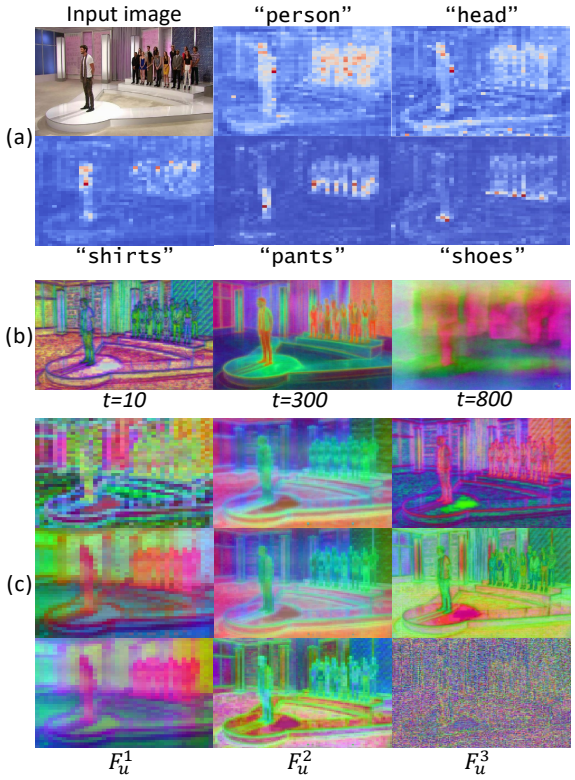

Figure 3. (a) Cross-attention maps highlighting different semantic regions based on textual queries. (b) PCA [42] visualization of features extracted from the ViT [12] block of $\mathbf{F}_u^3$ at different timesteps (t). (c) PCA visualization of up-stage feature maps ($\mathbf{F}_u^1, \mathbf{F}_u^2, \mathbf{F}_u^3$), with columns representing different up-stage levels and rows corresponding to features from ViT blocks within each level.

Timesteps. The effectiveness of diffusion features varies significantly depending on the timestep at which they are obtained, as shown in Fig. 3 (b). In datasets [9, 36] where objects are sparse and easily distinguishable, features extracted at early-to-middle timesteps (*e.g.,* t=300) tend to be the most informative. At very early timesteps (*e.g.,* t=10), the model removes only minimal noise, making feature extraction too trivial to capture meaningful representations. Conversely, at later timesteps (*e.g.,* t=800), excessive noise forces the model to focus more on denoising rather than preserving fine-grained details, leading to a degradation in feature quality [44, 75, 76]. However, this pattern may differ in the person search, which involves cluttered backgrounds, multiple objects, and diverse real-world conditions. These images exhibit high visual complexity even before artificial noise is introduced, making it harder to extract clear and discriminative features. Additionally, real-world noise from camera artifacts, motion blur, and lighting variations further intensifies the challenge. In such cases, adding synthetic noise exacerbates the difficulty of distinguishing meaningful features from irrelevant information. As a result, extracting features at earlier timesteps may be more beneficial (Fig. 3 (b)), as they preserve finer detailes while minimizing interference from both synthetic and real-world noise. This tendency further supported by our empirical results.

Hierarchical structure. The UNet [55] architecture in diffusion models follows a hierarchical structure (Fig. 2). In particular, the up-stage restores spatial information by integrating local features from the down-stage via skip connections while incorporating global context from the mid-stage. Consequently, up-stage features tend to be more informative than those from earlier stages, as they retain both fine-grained details and high-level semantics [40, 43, 75], as shown in Fig. 3 (c). In person search, as detection requires spatial precision and re-ID relies on distinctive features, the up-stage features may appear well-suited for both tasks. However, not all layers within the up-stage may be equally beneficial. Since the UNet is originally designed for noise prediction, its final layers (Fig. 3 (c), row 3, column 3) tend to focus on estimating noise maps rather than extracting meaningful features. Furthermore, even within the same hierarchical level, different layers exhibit distinct feature properties due to variations in convolutional operations, attention mechanisms, and residual connections. Our experiments confirm that a careful selection of specific up-stage layers significantly improves person search performance.

Shape bias. Pre-trained diffusion models [51, 52, 54, 56] have demonstrated superior discriminative power across various downstream tasks [6, 34, 40, 49, 70, 74] compared to traditional discriminative models [12, 19, 39]. This advantage stems from their ability to capture both global context and fine-grained features. Notably, diffusion features naturally emphasize global structures due to their progressive denoising process, which reconstructs images from noise by first recovering low-frequency components before refining finer details. While this property enhances robustness and improves high-level feature alignment, it may also introduce a shape bias, where low-frequency components tend to be more dominant than fine textures [24, 29]. For person search, where both global structure and fine-grained details are essential, leveraging diffusion features while further enhancing high-frequency information can lead to even greater discriminative power [15, 65].

3.3. Framework Overview

The overall architecture of our DiffPS is illustrated in Fig. 4 (a). DiffPS comprises three main components: a frozen pre-trained diffusion model backbone and two branches for detection and re-ID. Through empirical analysis (Table 3), we found that the detection branch can be effectively achieved using only a single feature map and a cross-attention map, while the re-ID branch benefits from leveraging multiple feature maps to capture discriminative details. Both branches are carefully designed to maximize the prior knowledge embedded in the pre-trained diffu-

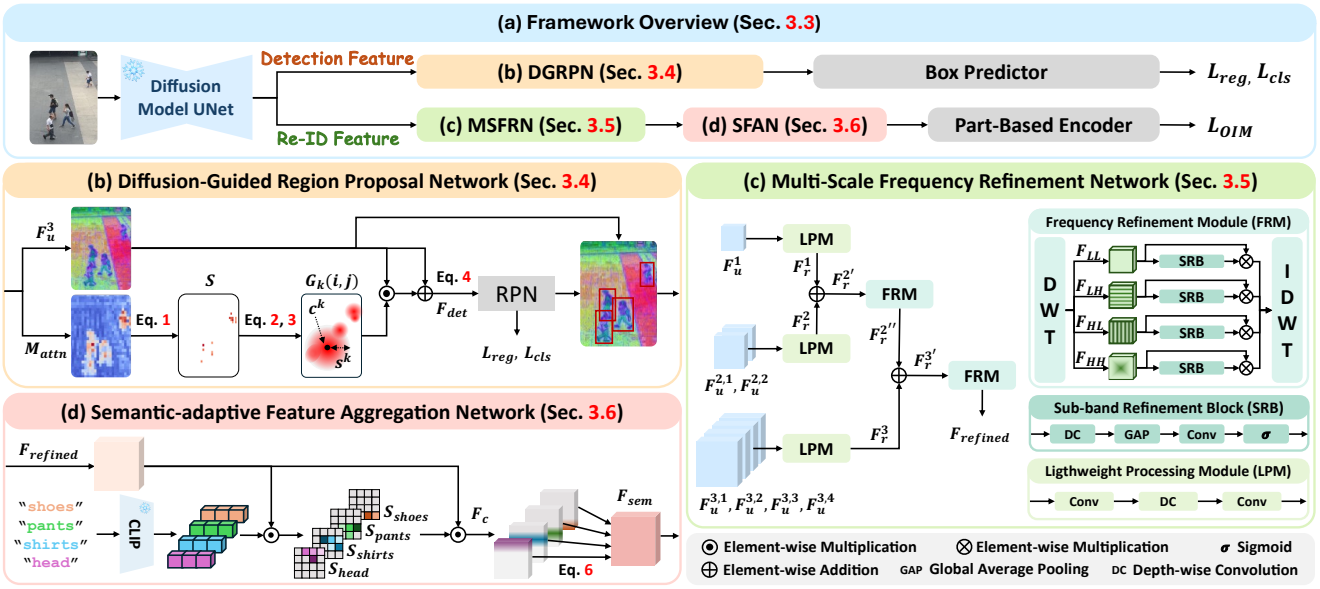

Figure 4. (a) Overview of DiffPS framework. DiffPS leverages a pre-trained diffusion model's UNet as the backbone, with three specialized modules: (b) DGRPN refines region proposals using cross-attention maps, (c) MSFRN enhances high-frequency details via multi-scale frequency refinement, and (d) SFAN incorporates text-aligned semantic features for re-ID. (Best viewed in color.)

sion model. For the detection branch, we adopt Faster R-CNN [53], following previous methods [17, 23, 26, 30, 33, 45, 69, 73]. However, instead of a conventional Region Proposal Network (RPN) [53], we introduce a Diffusion-Guided Region Proposal Network (DGRPN), which utilizes diffusion priors to guide the localization of potential person regions. In the re-ID branch, we first refine the diffusion features using a Multi-Scale Frequency Refinement Network (MSFRN) to enhance high-frequency details which are critical for identity discrimination. The refined features are then processed via RoI-Align [53], and we exploit the strong alignment between diffusion features and text embeddings by incorporating Semantic-adaptive feature aggregation network (SFAN). We further adopt a simple stripe-based partitioning network, which is widely used in re-ID methods [13, 26, 60, 61, 64], to extract the final person representation. Our decoupled structure allows for flexible architecture choices, enabling any detection or re-ID module to be seamlessly integrated in a plug-and-play manner, as demonstrated in the Suppl. B. For training, we employ standard loss functions: Smooth-L1 loss and cross-entropy loss for detection, and Online Instance Matching (OIM) [69] loss for re-ID. In the following sections, we provide detailed explanations of each module in our framework.

3.4. Diffusion-Guided Region Proposal Network

DGRPN leverages the strong alignment between diffusion features and text embeddings by utilizing the cross-attention mechanism associated with the `"person"` token embedding, as discussed in Sec. 3.2. Specifically, we extract the feature map $\mathbf{F}_u^3$ and the cross-attention map $\mathbf{M}_{att}$ from the corresponding cross-attention layer. The cross-attention map encodes attention scores that highlight person-related regions. However, due to the complexity of person search datasets with their cluttered backgrounds and multiple overlapping individuals, this attention map may lack sharpness or be imprecise. To address this, we first threshold the cross-attention map to retain only high-confidence regions:

$$\mathbf{M}_{th}(i,j) = \begin{cases} \mathbf{M}_{att}(i,j), & \text{if } \mathbf{M}_{att}(i,j) > \tau \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where τ is a predefined threshold. Next, we use the set of sampled pixels $\mathcal{S} = \{(i,j) \mid \mathbf{M}_{th}(i,j) > 0\}$ as Gaussian centers. Each selected pixel $(i,j) \in \mathcal{S}$ serves as a candidate center (c_x^k, c_y^k) for a Gaussian distribution, where k indexes different detected peaks. For each sampled pixel, we define its Gaussian standard deviation (s_w^k, s_h^k) based on local spatial statistics:

$$s_w^k = \max\left(\delta, \sqrt{\sum_{i,j \in \mathcal{N}(c_x^k, c_y^k)} (i - c_x^k)^2 \cdot \mathbf{M}_{th}(i,j)}\right), \tag{2}$$

where $\mathcal{N}(c_x^k, c_y^k)$ represents a local neighborhood around (c_x^k, c_y^k), and δ is a hyperparameter ensuring a minimum size for the Gaussian. The standard deviation s_h^k is computed in the same manner for the vertical axis. Using these centers and variances, we construct Gaussian maps:

$$G_k(i,j) = \exp\left(-\frac{(i - c_x^k)^2}{\beta (s_w^k)^2} - \frac{(j - c_y^k)^2}{\beta (s_h^k)^2}\right), \tag{3}$$

where β is a learnable scaling factor. Since multiple peaks can exist in the attention map, we aggregate all Gaussian maps by taking the element-wise maximum across them, producing the final map $G_{\text{det}}(i,j) = \max_k G_k(i,j)$. Finally, we modulate the detection-specific feature map $\mathbf{F}_u^3$ using the generated Gaussian map as follows:

$$\mathbf{F}_{\text{det}} = \mathbf{F}_u^3 + \gamma \left(G_{\text{det}} \odot \mathbf{F}_u^3 \right), \tag{4}$$

where $\odot$ denotes element-wise multiplication, and γ is a learnable parameter that controls the influence of the Gaussian-modulated features. This enhanced feature map $\mathbf{F}_{\text{det}}$ is used as input to the region proposal network.

3.5. Multi-Scale Frequency Refinement Network

As discussed in Sec. 3.2, diffusion features effectively capture both global context and fine-grained details. However, they tend to exhibit a shape bias due to the denoising process. To mitigate this, we propose MSFRN to improve discriminability by enhancing high-frequency details in diffusion features. Specifically, we first extract re-ID specific multi-scale feature maps from the up-stage, using four feature maps from $\mathbf{F}_u^3$, two from $\mathbf{F}_u^2$, and one from $\mathbf{F}_u^1$, yielding a total of seven feature maps. Then, MSFRN processes each feature independently to capture unique characteristics. To be specific, for each feature map $\mathbf{F}_u^{l,i}$, where i indexes different layers within the same level l, we apply a Lightweight Processing Module (LPM) consisting of 1×1 and depth-wise convolutions: $\mathbf{F}_r^{l,i} = \text{Conv}_{1\times1}\left(\text{DepthConv}\left(\text{Conv}_{1\times1}(\mathbf{F}_u^{l,i})\right)\right)$. These processed features are then aggregated within each level using concatenation followed by a 1×1 convolution, yielding three multi-scale refined features $\mathbf{F}_r^3, \mathbf{F}_r^2$, and $\mathbf{F}_r^1$.

To enhance high-frequency details, we adopt a hierarchical frequency decomposition strategy within the Frequency Refinement Module (FRM). The lowest-resolution feature $\mathbf{F}_r^1$ is first upsampled and added to $\mathbf{F}_r^2$, followed by Layer Normalization to yield $\mathbf{F}_r^{2'}$. We then apply the Discrete Wavelet Transform (DWT) to $\mathbf{F}_r^{2'}$, obtaining four sub-bands: a low-frequency component $\mathbf{F}_{LL}$ and three high-frequency components $\mathbf{F}_{LH}, \mathbf{F}_{HL}, \mathbf{F}_{HH}$. Each high-frequency band is refined using a Sub-band Refinement Block (SRB), consisting of depth-wise separable convolution, global average pooling, a 1×1 convolution, and a sigmoid activation to produce a channel-wise attention vector $\mathbf{s}_X$. This vector modulates the feature as: $\hat{\mathbf{F}}_X = \mathbf{s}_X \cdot \text{DepthConv}(\mathbf{F}_X), \quad X \in \{LH, HL, HH\}$. To reconstruct the refined feature, we apply the Inverse DWT with learnable scaling factors γ_X:

$$\mathbf{F}_r^{2''} = \text{IDWT}(\mathbf{F}_{LL}, \gamma_{LH}\hat{\mathbf{F}}_{LH}, \gamma_{HL}\hat{\mathbf{F}}_{HL}, \gamma_{HH}\hat{\mathbf{F}}_{HH}). \tag{5}$$

This output is upsampled and added to $\mathbf{F}_r^3$, where the same process is repeated. Finally, a 1×1 convolution is applied to the refined feature $\mathbf{F}_r^{3''}$ to obtain the final output $\mathbf{F}_{\text{refined}}$.

Our MSFRN effectively enhances fine-grained discriminative information by leveraging hierarchical frequency decomposition and multi-scale fusion, mitigating the loss of high-frequency details in shape-biased features.

3.6. Semantic-Adaptive Feature Aggregation Network

SFAN harnesses the strong alignment between diffusion features and text embeddings in their shared semantic space to enhance person representations. Given the refined feature map $\mathbf{F}_{\text{refined}}$ obtained from MSFRN, we first compute the similarity between each spatial feature and pre-defined text embeddings corresponding to human body regions, including `"head"`, `"shirts"`, `"pants"`, and `"shoes"`. Through experiments, we identify the optimal combination of text prompts, with additional details provided in the Suppl. C. These text embeddings, extracted from the CLIP [50] text encoder of a text-to-image diffusion model [52, 54], are computed only once. The cosine similarity between the refined feature map and each text embedding token is computed, producing four spatial semantic maps, $\mathbf{S}_{\text{head}}, \mathbf{S}_{\text{shirts}}, \mathbf{S}_{\text{pants}}, \mathbf{S}_{\text{shoes}}$, where each $\mathbf{S}_c$ highlights the relevance of each feature location to the corresponding body part c. We utilize these semantic maps to generate spatially adaptive person feature maps. Specifically, each semantic map $\mathbf{S}_c$ is normalized via a softmax function over all categories to obtain a probability distribution $\hat{\mathbf{S}}_c$. The refined feature map is then weighted by these normalized maps to emphasize semantic regions while preserving global spatial structure. This process is formulated as $\mathbf{F}_c = \hat{\mathbf{S}}_c \odot \mathbf{F}_{\text{refined}}$, where $\odot$ denotes element-wise multiplication, applying the semantic relevance as a spatial attention mechanism. To integrate all semantic regions while maintaining spatial consistency, the refined body-part features are aggregated as:

$$\mathbf{F}_{\text{sem}} = \sum_c \mathbf{W}_c \mathbf{F}_c, \tag{6}$$

where $\mathbf{W}_c$ is a learnable weight parameter that adaptively balances contributions from different body regions. Through this adaptive weighting mechanism, SFAN emphasizes informative body regions while suppressing irrelevant background noise and occluded regions.

4. Experiments

4.1. Experimental details

Datasets and evaluation metrics. Following prior methods [3, 5, 17, 33, 69, 72, 73], we evaluate our model on CUHK-SYSU [69] and PRW [77]. CUHK-SYSU contains 18,184 images, 96,143 pedestrian boxes, and 8,432 identities, with 11,206 images for training and 6,978 for testing. PRW consists of 11,816 frames from six cameras, with 43,110 boxes for 932 identities, including 5,704 training

Method	Backbone	CUHK-SYSU		PRW	
		mAP	Top-1	mAP	Top-1
OIM [69]	ResNet50	75.5	78.7	21.3	49.4
IAN [68]	ResNet50	76.3	80.1	23.0	61.9
QEEPS [45]	ResNet50	88.9	89.1	37.1	76.7
BINet [10]	ResNet50	90.0	90.7	45.3	81.7
APNet [79]	ResNet50	88.9	89.3	41.9	81.4
NAE [5]	ResNet50	91.5	92.4	43.3	80.9
NAE+ [5]	ResNet50	92.1	92.9	44.0	81.1
PGSFL [27]	ResNet50	90.2	91.8	42.5	83.5
SeqNet [33]	ResNet50	93.8	94.6	46.7	83.4
DMRNet [17]	ResNet50	93.2	94.2	46.9	83.3
AlignPS [72]	ResNet50	93.1	93.4	45.9	81.9
COAT [73]	ResNet50	94.2	94.7	53.3	87.4
PSTR [3]	ResNet50	93.5	95.0	49.5	87.8
PSTR [3]	PVTv2-B2	95.2	96.2	56.5	89.7
SeqNeXt [23]	ConvNeXt	96.1	96.5	57.6	89.5
SeqNeXt+GFN [23]	ResNet50	94.7	95.3	51.3	90.6
SeqNeXt+GFN [23]	ConvNeXt	96.4	97.0	58.3	**92.4**
SOLIDER [8]	Swin-S	95.5	95.8	59.8	86.7
SEAS [26]	ResNet50	96.2	97.1	52.0	85.7
SEAS [26]	ConvNeXt	97.1	97.8	60.5	89.5
DiffPS (Ours)	SD v2-1	**97.8**	**98.4**	**62.0**	91.0

Table 1. Comparison with the state-of-the-art methods on CUHK-SYSU [69] and PRW [77] test sets. Numbers in bold indicate the best performance and underscored ones are the second best.

and 6,112 test images. Performance is measured using mean average precision (mAP) and top-1 accuracy (Top-1).

Implementation details. We train our model end-to-end for 20 epochs. We use the Adam [28] optimizer, where β_1 and β_2 are set to 0.9 and 0.999, respectively. We use a warm-up and step decay strategy, linearly increasing the learning rate from a starting point of 1×10^{-7} to 1×10^{-4} over the first epoch. The batch size is set to 5 and random horizontal flip is used as the augmentation method. In DGRPN, the threshold τ is set to 0.7 and δ to 5. We use Stable Diffusion [54] v2-1 as the backbone. For the part-based encoder, we follow the design of SEAS [26].

4.2. Comparison to the state-of-the-arts

CUHK-SYSU. On the CUHK-SYSU [69] dataset, our DiffPS achieves state-of-the-art performance with 97.8% mAP and 98.4% Top-1 accuracy, surpassing all previous methods. Notably, DiffPS outperforms ResNet50 [19]-based methods by a large margin of 1.6-22.3% in mAP. Even when compared to methods using more advanced backbones like ConvNeXt [39] and Swin-S [38], DiffPS demonstrates superior performance, highlighting the effectiveness of leveraging the diffusion priors for person search.

PRW. The PRW dataset presents a more challenging scenario, yet DiffPS maintains its superior performance with 62.0% mAP and 91.0% Top-1 accuracy. This represents a significant improvement of 2.2% mAP over SOLIDER [8] and 1.5% over SEAS [26] with ConvNeXt [39]. While our Top-1 accuracy is slightly lower than SeqNeXt+GFN [23], we achieve a significantly higher mAP.

Method	Detection	Re-ID	
	AP	mAP	Top-1
DMRNet [17] (joint optimized)	86.6(**-1.5**)	93.2(**-0.4**)	94.2(**-0.8**)
DMRNet [17] (detection only)	88.1	-	-
DMRNet [17] (re-ID only)	-	93.6	95.0
DMRNet++ [18] (joint optimized)	88.3 (**-1.1**)	94.4 (**-0.7**)	95.5 (**-1.1**)
DMRNet++ [18] (detection only)	89.4	-	-
DMRNet++ [18] (re-ID only)	-	95.1	96.6
Ours (joint optimized)	90.9	97.8	98.4
Ours (detection only)	90.9	-	-
Ours (re-ID only)	-	97.8	98.4

Table 2. Comparison of joint and individual task optimization on CUHK-SYSU [69] test set. Numbers in parentheses indicate performance drop from individual to joint optimization.

Type	Combination	mAP	Top-1
(a)	Level 1 Only	50.0	85.6
(b)	Level 2 Only	58.0	89.4
(c)	Level 3 Only	56.7	89.0
(d)	Level 0 + 2	56.1	86.9
(e)	Level 1 + 2	60.3	89.7
(f)	Level 2 + 3	61.3	**91.2**
(g)	**Level 1 + 2 + 3**	**62.0**	91.0
(h)	All levels combined	60.5	90.8

Table 3. Ablation study on feature map selection for re-ID on PRW [77] test set. (a)-(c) use the top 7 highest performing features from a single level, (d)-(f) utilize 4 from each level, (g) combines 1, 2, and 4, and (h) uses 2 from each level.

4.3. Ablation Study

Resolving conflict. As shown in Table 2, we compare our method with DMRNet [17] and DMRNet++ [18], which also employ a decoupled design but still suffers from optimization conflicts. These methods exhibits a noticeable performance degradation when jointly optimizing detection and re-ID tasks compared to their individual optimization. This degradation highlights the challenge of using a shared backbone, where competing gradients from different tasks interfere with each other during training. Despite its decoupled design, they cannot fully eliminate this conflict as it updates the shared backbone with contradictory objectives. In contrast, our method achieves identical performance whether trained jointly or separately for each task. This perfect decoupling stems from maximizing diffusion priors through specialized modules without the need to update the shared backbone.

Feature selection. Selecting task-specific features for detection and re-ID requires careful consideration. While detection achieves strong performance with a single feature map, re-ID benefits from a thoughtful selection of multiple feature maps to capture richer discriminative information. Table. 3 presents the impact of different feature selection strategies on re-ID performance. This experiment is based on the layer-wise analysis provided in the Suppl. J. The results demonstrate that combining multi-scale features from multiple levels consistently outperforms single-level

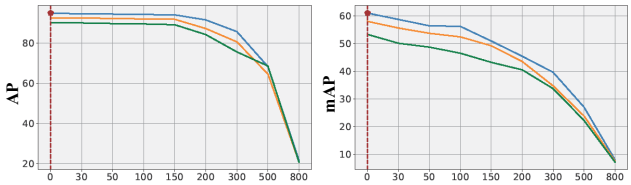

Figure 5. Timestep analysis of diffusion features on PRW [77]. (a) Detection performance (AP) using three different layers from upstage level 3 (blue, orange, green). (b) Re-ID performance (mAP) using features corresponding to (g), (b), and (c) from Table 3 (blue, orange, green). (Best viewed in color.)

features. This may be attributed to the complementary information captured at different scales. Notably, level 0 features performed poorly due to their limited spatial resolution, which restricts their capacity to encode discriminative features. The optimal performance is achieved with (g), yielding 62.0% mAP and 91.0% Top-1 accuracy, which we adopt in our final configuration.

Timesteps for detection and re-ID. As shown in Fig. 5, we observe distinct performance trends across different timesteps. For detection (Fig. 5(a)), AP remains stable up to moderate timesteps but starts to decline beyond t=200. This suggests that detection benefits from noise acting as a regularizer, improving robustness to varying scene conditions. In contrast, re-ID performance (mAP) deteriorates more rapidly, with a monotonic decline (Fig. 5(b)). Since re-ID relies on fine-grained details for identity discrimination, even small amounts of noise disrupt feature consistency, making it more sensitive to timestep variations. Notably, both tasks achieve optimal performance at t=0, where the features preserve clear semantic structure without noise interference, supporting the analysis in Sec. 3.2.

DGRPN. We conduct an ablation study to evaluate the effectiveness of DGPRN, as shown in Table 4. The baseline is Faster R-CNN [53], a widely used in prior methods [23, 26, 33, 69], and we observe that incorporating DGPRN improves performance. This indicates that diffusion priors effectively guide region proposals by highlighting potential areas of interest. In the ablation study for τ, the results show that when τ is too low, unnecessary regions are included, causing a drop in AP, while a high τ removes valuable information, reducing recall. The optimal performance is achieved at $\tau = 0.5$, balancing precision and recall.

MSFRN and SFAN. Table 5 presents the ablation study on PRW [77], evaluating the impact of MSFRN and SFAN. The baseline employs DGRPN for detection and a part-based encoder for re-ID. Integrating MSFRN enhances mAP (+2.5%) and Top-1 accuracy (+2.5%) by refining high-frequency details, which are crucial for distinguishing similar individuals through identity-discriminative features. As shown in Fig. 6 (a), MSFRN effectively amplifies fine-grained patterns and high frequency information, as evident in the refined feature maps and corresponding high-frequency components. Adding SFAN further improves mAP (+0.5%) and Top-1 accuracy (+0.3%) by leveraging text-aligned semantic features, which enhance robustness to occlusions and pose variations through region-aware representations. Fig. 6 (b) illustrates how the semantic representation $\hat{\mathbf{S}}_c$ highlights discriminative body parts, while the final refined feature $\mathbf{F}_{\text{sem}}$ effectively suppresses background noise and occlusions, ensuring a stronger focus on the person. Further experiment on occlusion robustness is provided in the Suppl. F. Combining MSFRN and SFAN in DiffPS yields the best performance (62.0% mAP, 91.0% Top-1), confirming that the two modules complement each other to improve re-ID performance.

Method	AP	Recall
Baseline	94.2	97.5
+ DGRPN	**94.8**	**98.1**
$\tau = 0.2$	94.4	97.9
$\tau = 0.5$	**94.8**	**98.1**
$\tau = 0.8$	94.7	97.6

Table 4. Ablation study of DGRPN on PRW [77].

Components		Performance	
MSFRN	SFAN	mAP	Top-1
		59.1	88.1
✓		61.6	90.6
	✓	59.6	88.4
✓	✓	**62.0**	**91.0**

Table 5. Ablation study of MSFRN and SFAN on PRW [77].

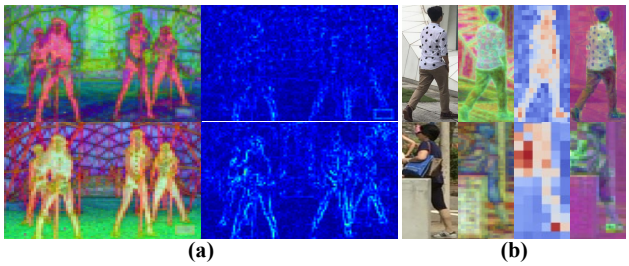

Figure 6. Qualitative results of MSFRN and SFAN. (a) The first row shows features before MSFRN, while the second row presents the refined outputs, with the right column visualizing high-frequency components via DWT. (b) The first column displays original person crops, and the second column presents PCA [42] visualizations of features before being processed by SFAN. The third column shows the aggregated semantic maps $\sum_c \hat{\mathbf{S}}_c$, and the last presents $\mathbf{F}_{\text{sem}}$, the final semantic features.

5. Conclusion

In this paper, we propose DiffPS, a novel person search framework that fully leverages the prior knowledge of pre-trained diffusion models to address the fundamental challenges of backbone generalization and task conflict. Based on our analysis of diffusion priors, we introduce three specialized modules that maximize their effectiveness for person search. DiffPS completely resolves the optimization conflict between detection and re-ID while achieving state-of-the-art performance on CUHK-SYSU and PRW.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00355008). This research was also supported by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Developing Professionals for R&D in Contents Production Based on Generative AI and Cloud, Project Number: RS-2024-00352578, Contribution Rate: 25%). Additionally, this work was supported by the National Research Foundation (NRF) grant (RS-2024-00421203), and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)).

References

[1] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. The eurocity persons dataset: A novel benchmark for object detection. *arXiv preprint arXiv:1805.07193*, 2018. 3

[2] Jiale Cao, Yanwei Pang, Jin Xie, Fahad Shahbaz Khan, and Ling Shao. From handcrafted to deep features for pedestrian detection: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4913–4934, 2021. 1

[3] Jiale Cao, Yanwei Pang, Rao Muhammad Anwer, Hisham Cholakkal, Jin Xie, Mubarak Shah, and Fahad Shahbaz Khan. Pstr: End-to-end one-step person search with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9458–9467, 2022. 1, 2, 3, 6, 7

[4] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *Proceedings of the european conference on computer vision (ECCV)*, pages 734–750, 2018. 1, 2

[5] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12615–12624, 2020. 1, 2, 3, 6, 7

[6] Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Your diffusion model is secretly a certifiably robust classifier. *arXiv preprint arXiv:2402.02316*, 2024. 2, 3, 4

[7] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19830–19843, 2023. 3

[8] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15050–15061, 2023. 3, 7

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2, 4

[10] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Bi-directional interaction network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2839–2848, 2020. 7

[11] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Instance guided proposal network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2585–2594, 2020. 1, 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4

[13] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. *Advances in neural information processing systems*, 32, 2019. 5

[14] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14750–14759, 2021. 3

[15] Tianyu Zhang Bo Li1 Shiliang Pu Guiwei Zhang, Yongfei Zhang. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14133–14142, 2023. 4

[16] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9814–9823, 2019. 1, 2

[17] Chuchu Han, Zhedong Zheng, Changxin Gao, Nong Sang, and Yi Yang. Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1505–1512, 2021. 1, 2, 3, 5, 6, 7

[18] Chuchu Han, Zhedong Zheng, Kai Su, Dongdong Yu, Zehuan Yuan, Changxin Gao, Nong Sang, and Yi Yang. Dmrnet++: Learning discriminative features with decoupled networks and enriched pairs for one-step person search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7319–7337, 2023. 7

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2, 4, 7

[20] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF inter-*

national conference on computer vision, pages 4918–4927, 2019. 2

[21] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 1

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[23] Lucas Jaffe and Avideh Zakhor. Gallery filter network for person search. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1684–1693, 2023. 1, 2, 3, 5, 7, 8

[24] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779*, 2023. 4

[25] Chengyou Jia, Minnan Luo, Zhuohang Dang, Guang Dai, Xiaojun Chang, and Jingdong Wang. Psdiff: Diffusion model for person search with iterative and collaborative refinement, 2024. 3

[26] Yimin Jiang, Huibing Wang, Jinjia Peng, Xianping Fu, and Yang Wang. Scene-adaptive person search via bilateral modulations. *arXiv preprint arXiv:2405.02834*, 2024. 1, 2, 3, 5, 7, 8

[27] Hanjae Kim, Sunghun Joung, Ig-Jae Kim, and Kwanghoon Sohn. Prototype-guided saliency feature learning for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4865–4874, 2021. 7

[28] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[29] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Text-to-image diffusion models are great sketch-photo matchmakers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16826–16837, 2024. 4

[30] Sanghoon Lee, Youngmin Oh, Donghyeon Baek, Junghyup Lee, and Bumsub Ham. Oimnet++: Prototypical normalization and localization-aware learning for person search. In *European Conference on Computer Vision*, pages 621–637. Springer, 2022. 5

[31] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023. 2, 3

[32] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 3

[33] Zhengjia Li and Duoqian Miao. Sequential end-to-end network for efficient person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2011–2019, 2021. 1, 2, 3, 5, 6, 7, 8

[34] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7667–7676, 2023. 2, 3, 4

[35] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015. 1

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 4

[37] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5187–5196, 2019. 1

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[39] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 1, 2, 4, 7

[40] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4

[41] Chaofan Ma, Yuhuan Yang, Chen Ju, Fei Zhang, Jinxiang Liu, Yu Wang, Ya Zhang, and Yanfeng Wang. Diffusionseg: Adapting diffusion towards unsupervised object discovery. *arXiv preprint arXiv:2303.09813*, 2023. 2, 3

[42] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19 (3):303–342, 1993. 1, 4, 8

[43] Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, and Qingming Huang. Not all diffusion model activations have been evaluated as discriminative features. *arXiv preprint arXiv:2410.03558*, 2024. 2, 3, 4

[44] Soumik Mukhopadhyay, Matthew Gwilliam, Yosuke Yamaguchi, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Tianyi Zhou, Jun Ohya, and Abhinav Shrivastava. Do text-free diffusion models learn discriminative visual representations? In *European Conference on Computer Vision*, pages 253–272. Springer, 2024. 4

[45] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2019. 5, 7

[46] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[47] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3

[48] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4967–4975, 2019. 1

[49] Koutilya Pnvr, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, and David Jacobs. Ld-znet: A latent diffusion approach for text-based image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4157–4168, 2023. 2, 3, 4

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[51] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2, 3, 4

[52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2, 3, 4, 6

[53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 5, 8

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4, 6, 7

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3, 4

[56] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3, 4

[57] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2, 3

[58] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 3

[59] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[60] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3960–3969, 2017. 5

[61] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 5

[62] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3554–3563, 2024. 2, 3

[63] Yanling Tian, Di Chen, Yunan Liu, Jian Yang, and Shanshan Zhang. Divide and conquer: Hybrid pre-training for person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5224–5232, 2024. 2, 3

[64] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, page 274–282. ACM, 2018. 1, 5

[65] Jingye Wang, Ruoyi Du, Dongliang Chang, Kongming Liang, and Zhanyu Ma. Domain generalization via frequency-domain-based feature disentanglement and interaction. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4821–4829, 2022. 4

[66] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 3

[67] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1206–1217, 2023. 2, 3

[68] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. Ian: the individual aggregation network for person search. *Pattern Recognition*, 87: 332–340, 2019. 7

[69] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3415–3424, 2017. 1, 2, 3, 5, 6, 7, 8

[70] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2, 3, 4

[71] Zhengrui Xu, Guan'an Wang, Xiaowen Huang, and Jitao Sang. Denoiserep: Denoising model for representation learning, 2024. 3

[72] Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. Anchor-free person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2021. 1, 2, 3, 6, 7

[73] Rui Yu, Dawei Du, Rodney LaLonde, Daniel Davila, Christopher Funk, Anthony Hoogs, and Brian Clipp. Cascade transformers for end-to-end person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7267–7276, 2022. 1, 2, 3, 5, 6, 7

[74] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4

[75] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023. 3, 4

[76] Yuzhong Zhao, Qixiang Ye, Weijia Wu, Chunhua Shen, and Fang Wan. Generative prompt model for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6351–6361, 2023. 3, 4

[77] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017. 2, 6, 7, 8

[78] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017. 1

[79] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust partial matching for person search in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6827–6835, 2020. 7