Leveraging Prior Knowledge of Diffusion Model for Person Search

Supplementary Material

Detection Branch	Re-ID Branch	Detection		Re-ID	
		Recall	AP	mAP	Top-1
Faster R-CNN [15]	Ours	97.6	94.5	61.8	<u>90.8</u>
RetinaNet [10]		<u>97.8</u>	<u>94.6</u>	<u>61.9</u>	**91.0**
Ours	MGN [19]	**98.1**	**94.8**	59.1	88.4
	PCB [18]	**98.1**	**94.8**	60.4	90.1
	NAE [3]	**98.1**	**94.8**	60.0	89.1
	SEAS [8]	**98.1**	**94.8**	60.7	89.3
Ours	Ours	**98.1**	**94.8**	**62.0**	**91.0**

Table 1. Performance comparison of different detection and re-ID models on PRW [23] dataset. Numbers in bold indicate the best performance and underscored ones are the second best.

A. UNet Architecture in Diffusion Models

The UNet [17] architecture in diffusion models follows a hierarchical structure, consisting of three primary stages: down-stage, mid-stage, and up-stage. Each of these stages is composed of multiple resolution levels, where feature activations at the same resolution are processed by a series of specialized modules, including ResNet [6] blocks (Res blocks), Vision Transformer [4] blocks (ViT blocks), and up/down-samplers. These modules facilitate hierarchical feature extraction and enable efficient denoising by progressively reducing and restoring spatial resolution. The down-stage is responsible for reducing the spatial resolution of feature activations while increasing their channel depth. This stage comprises four resolution levels, with each level containing a sequence of Res blocks, ViT blocks, and down-samplers. The hierarchical nature of this stage allows the model to capture low-level details in the early layers and progressively extract more abstract and high-level features as the resolution decreases. At the lowest resolution, the mid-stage acts as a bottleneck layer that connects the down-stage and up-stage. It consists of stacked Res and ViT blocks, enabling feature refinement before upsampling begins. The up-stage mirrors the down-stage by progressively restoring spatial resolution through a sequence of Res blocks, ViT blocks, and up-samplers. Skip connections are established between corresponding levels in the down-stage and up-stage, allowing the network to propagate fine-grained details and prevent information loss.

B. Plug-and-Play Compatibility

In Table 1, we demonstrate competitiveness of our proposed modules with other state-of-the-arts detection modules (Faster R-CNN [15] and RetinaNet [10]) and re-ID modules (MGN [19], PCB [18], NAE [3], and SEAS [8]).

Text Prompts	**mAP**	**Top-1**
`"head", "upper body", "lower body", "foot"`	61.5	90.5
`"face", "torso", "legs", "foot"`	<u>61.7</u>	<u>90.8</u>
`"head", "shirts", "pants", "shoes"` (Ours)	**62.0**	**91.0**

Table 2. Ablation study on different text prompts for SFAN on PRW [23]. Using clothing-related prompts (`"shirts"` and `"pants"`) provides more stable and distinctive cues, leading to the best re-ID performance. Numbers in bold indicate the best performance and underscored ones are the second best.

Our detection branch, guided by the proposed Diffusion-Guided Region Proposal Network (DGRPN), achieves the highest recall (98.1%) and AP (94.8%), outperforming Faster R-CNN (97.6%, 94.5%) and RetinaNet (97.8%, 94.6%). This highlights the effectiveness of DGRPN in enhancing person localization using cross-attention maps. Additionally, our re-ID branch consistently outperforms existing re-ID modules. While SEAS [8] achieves a mAP of 60.7% and Top-1 accuracy of 89.3%, our method further improves the performance to 62.0% mAP and 91.0% Top-1 accuracy, demonstrating the benefits of our proposed modules in re-ID task.

C. Text prompt

To investigate the impact of different text prompts used in Semantic-adaptive feature aggregation network (SFAN), we conduct an ablation study by varying the predefined body-region text embeddings, as shown in Table 2. We compare three sets of prompts: (1) `"head"`, `"upper body"`, `"lower body"`, and `"foot"`, (2) `"face"`, `"torso"`, `"legs"`, and `"foot"`, and (3) `"head"`, `"shirts"`, `"pants"`, and `"shoes"`. The results indicate that the third configuration achieves the best performance, with the highest mAP and Top-1 accuracy. This improvement is attributed to the fact that `"shirts"` and `"pants"` explicitly correspond to clothing attributes, which are more stable and visually distinctive compared to `"upper body"` or `"torso"`, which may introduce ambiguity due to pose variations and occlusions. Similarly, `"shoes"` provide a clearer distinction than `"foot"`, as they often contain more discriminative patterns (*e.g.*, color or style differences) that aid re-identification. In contrast, configurations (1) and (2) show degraded performance, likely due to their reliance on more generalized body descriptors that do not directly capture clothing details, leading to less discriminative spatial attention maps. These findings confirm that selecting text prompts that directly correspond to clothing-related features improves the effectiveness of SFAN in enhancing person representations.

Agg Net.	Re-ID	
	mAP	**Top-1**
Hyperfeature [11]	60.9	90.2
CWA [20]	60.6	90.8
Ours (MSFRN)	**62.0**	**91.0**

Table 3. Ablation study on various aggregation networks. Our proposed MSFRN achieves superior mAP and Top-1 accuracy. Numbers in bold indicate the best performance and underscored ones are the second best.

Backbone	Detection		Re-ID	
	Recall	AP	mAP	Top-1
DINO [13] ViT-B [4]	75.2	70.4	33.5	66.1
DINO [13] ViT-L [4]	81.3	76.5	36.1	72.8
DINO [13] ViT-G [4]	84.5	79.8	41.5	76.8
SD v1-5 [16]	97.8	**94.8**	61.3	89.7
SD v2-1 [16]	**98.1**	**94.8**	**62.0**	**91.0**

Table 4. Comparison of different pre-trained frozen backbones in our framework. We compare Stable Diffusion [14, 16] (SD) v1-5 and v2-1 with DINO [2, 13] models of varying sizes (Base, Large, Giant) on the PRW [23] dataset. Numbers in bold indicate the best performance and underscored ones are the second best.

D. Feature aggregation network

We investigate the impact of different aggregation network architectures on person search performance, as shown in Table 3. We compare our MSFRN against several existing networks, including Hyperfeature [11] (Res block-based) and CWA [20]. Our proposed MSFRN, consisting of a multi-scale frequency refinement strategy, achieves superior performance with 62.0% mAP and 91.0% Top-1 accuracy, outperforming existing methods. This improvement stems from MSFRN's ability to effectively preserve high-frequency details while maintaining global feature coherence, enabling the extraction of more discriminative identity representations.

E. Pre-trained Backbone Selection

In Table 4, we compare two different types of pre-trained foundation models as our backbone: DINO [13], trained via self-supervised learning, and Stable Diffusion (SD) [16], trained through text-to-image generative modeling. We compare against DINO considering its strong performance in various visual recognition tasks. For fair comparison, we carefully configure DINO's feature extraction: the last layer token features are used for detection to leverage high-level semantic understanding, while features from the last seven layers are aggregated for re-ID. Our results show that SD significantly outperforms DINO variants across all metrics. While DINO learns to align representations between teacher and student networks, SD learns to reconstruct the complete visual hierarchy through the denoising process. The iterative denoising process of SD enables the model to learn both fine-grained appearance details and global structural information simultaneously, which naturally aligns with both requirements of person search. This comprehensive feature learning proves more effective than the instance-level discrimination of DINO, as evidenced by superior detection performance and re-ID accuracy.

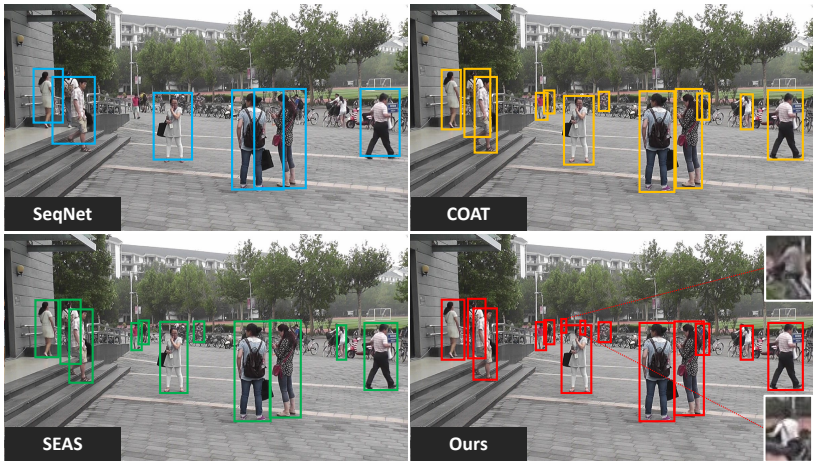
Method	Re-ID	
	mAP	Top-1
COAT†	86.5	85.6
SeqNeXt	91.1	89.8
SeqNeXt+GFN	92.0	90.9
SEAS†	89.6	87.7
Ours	**93.0**	**91.9**

Table 5. Occluded re-ID performance comparison across different methods on CUHK-SYSU [21]. Performance metrics using occluded person queries, demonstrating the effectiveness of our method under occlusion conditions. †: Methods directly implemented or reproduced by us.

F. Key Challenges in Person Search

Occluded person search. We show in Table 5 the robustness of our DiffPS to occlusion in person search. The evaluation protocol consists of 187 occluded person queries paired with a gallery of 50 images, where each query contains significant occlusion to simulate real-world scenarios. While occlusion poses a significant challenge in person search due to incomplete visual information, our framework achieves state-of-the-art performance (mAP = 93.0%, Top-1 = 91.9%) on the occluded person retrieval task. This superior performance under occlusion can be attributed to the generative nature of diffusion models, which learn to reconstruct complete visual information through the denoising process. This learned ability to recover missing or corrupted visual details enables our model to maintain robust person matching even when key body parts are occluded.

Small-scale person detection. Person search requires accurate person detection across various scales. While existing state-of-the-art methods [1, 7, 8, 22] achieve strong performance on medium and large-scale persons, detecting small-scale persons remains a significant challenge. We demonstrate in Fig. 1 our model's superior capability in addressing this challenge. We define small-scale instances as those whose bounding box areas fall within the bottom 25% of all bounding box areas in the dataset. As shown in the left of Fig. 1, our framework achieves superior performance in small object detection (AP_{small} = 94.7%) compared to existing methods. This strong performance on small instances may stems from two key characteristics of diffusion models: 1) the iterative denoising process inherently requires the model to learn multi-scale feature representations, from fine details to global structures, making it partic-



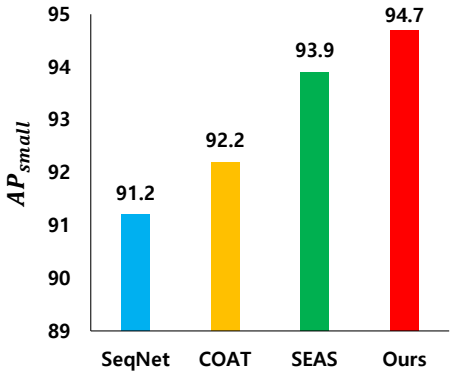


Figure 1. Qualitative and quantitative comparison of small person detection performance. Left: Quantitative comparison of AP_{small} scores on the PRW [23] test set across different methods, showing our model's superior performance in small person detection. Right: Visual comparison between SeqNet [9], COAT [22], SEAS [8], and our method on a challenging scene from the PRW test set containing multi-scale persons. Different colored boxes indicate detection results from each method.

	Method	Backbone	PRW		CUHK-SYSU	
			Detection	Re-ID	Detection	Re-ID
(a)	COAT	ResNet50	93.3 / 96.0	53.3 / 87.4	88.3 / 91.6	94.2 / 94.7
(b)	COAT	SD v2-1	94.1 / 96.3	58.9 / 89.5	89.5 / 92.9	95.3 / 96.1
(c)	SEAS	ConvNeXt	94.3 / 97.6	60.5 / 89.5	90.0 / 93.6	97.1 / 97.8
(d)	SEAS	SD v2-1	94.5 / 97.5	60.8 / 90.1	90.3 / 93.9	97.3 / 97.7
(e)	Baseline (B)	SD v2-1	94.2 / 97.5	59.1 / 88.1	90.2 / 94.0	95.5 / 96.1
(f)	B + D	SD v2-1	**94.8 / 98.1**	59.2 / 88.3	**90.9 / 94.4**	95.6 / 96.2
(g)	B + D + S	SD v2-1	**94.8 / 98.1**	59.6 / 88.5	**90.9 / 94.4**	96.4 / 96.8
(h)	B + D + M	SD v2-1	**94.8 / 98.1**	61.6 / 90.6	**90.9 / 94.4**	97.0 / 97.8
(i)	B + D + M + S	SD v2-1	**94.8 / 98.1**	**62.0 / 91.0**	**90.9 / 94.4**	**97.8 / 98.4**

Table 6. D: DGRPN, M: MSFRN, S: SFAN. Detection is evaluated by AP / Recall, and Re-ID by mAP / Top-1.

ularly effective at capturing small object features; 2) diffusion models are trained on large-scale datasets with diverse scene compositions, enabling them to learn robust representations of objects at various scales and contexts. The qualitative comparison in the right of Fig. 1 clearly demonstrates this advantage. In a challenging scene with multiple small persons against a cluttered background, our proposed DiffPS demonstrates superior detection performance on small-scale persons compared to existing methods. This visual evidence indicates that the prior knowledge learned through generative modeling is particularly beneficial for challenging scenarios like small object detection, even without task-specific fine-tuning.

G. Module Effectiveness

To rigorously validate the effectiveness of our proposed modules beyond the impact of the backbone itself, we conduct experiments using the same diffusion backbone across existing methods, as shown in Table 6. Specifically, rows (b), (d), and (i) demonstrate that even when applying SD v2-1 to existing frameworks, our method still achieves superior performance. This indicates that our performance gains are not simply due to the choice of a stronger backbone. Furthermore, row (e) presents a baseline that utilizes the SD v2-1 backbone without any of our proposed modules. Notably, this baseline performs worse than existing methods, highlighting that the backbone alone is insufficient to achieve state-of-the-art performance. From rows (e) to (i), we incorporate our proposed modules into the baseline, clearly showing that each module contributes meaningfully to performance improvement.

H. Shape Bias

To directly validate that MSFRN mitigates shape bias, we conduct an experiment using the Cue-conflict [5] dataset, which is specifically designed to test whether a model relies more on shape or texture. As shown in Fig. 2, this dataset contains images where the shape belongs to one class, but the texture is replaced with that of a different class. If the model predicts the label based on the shape,

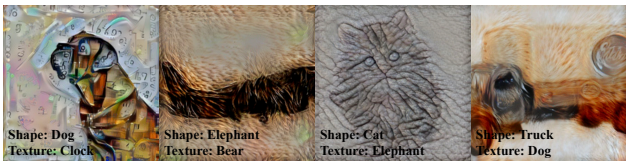


Figure 2. Cue-conflict examples with shape/texture labels and model prediction probabilities with and without MSFRN.

Model	Shape ↓
ResNet50	28.18
+ MSFRN	**26.32**
SD v2-1	63.52
+ MSFRN	**58.28**

Table 7. Shape bias mitigation.

δ	AP	Recall
1	94.3	97.6
3	94.7	97.9
5	**94.8**	**98.1**
7	94.6	97.7

Table 8. Ablation study on δ

it means the model is biased toward shape information. For example, in the first image of Fig. 2, where the shape corresponds to a dog and the texture to a clock, a shape-biased model would classify it as a dog. Table 7 and Fig. 2 show the shape classification accuracy with and without MSFRN. Applying MSFRN reduces shape bias in both models, suggesting its effectiveness in reducing shape reliance and enhancing focus on fine-grained textures.

I. Effect of δ.

We investigate the effect of the hyperparameter δ in our Gaussian proposal mechanism within the Diffusion-Guided Region Proposal Network (DGRPN). δ controls the minimum spatial extent of the Gaussian standard deviation used to modulate attention-based proposals. As shown in our ablation study on the PRW dataset, both overly small and large δ values degrade performance: small values fail to suppress noisy or irrelevant regions, while large values over-smooth the localization map, reducing precision. The best performance is achieved at $\delta = 5$, which effectively balances precision and recall, leading to optimal detection performance.

J. Analysis on Feature Map

Layer-wise analysis We demonstrate feature characteristics of different layers and modules within the UNet [17] architecture through quantitative and qualitative analysis. As shown in Tables 9 and 10, the up-stage features consistently outperform their down-stage and mid-stage counterparts across all metrics. While down-stage features show moderate performance and mid-stage features demonstrate notably degraded performance, up-stage features exhibit remarkably superior performance, particularly in levels 2 and 3. The superior performance of up-stage features is further validated through qualitative analysis, which also reveals how different modules at the same level complement each other. Figure. 3 shows that up-stage features from ResNet [6] (Res) modules, especially at levels 2 and 3, maintain more distinctive patterns than their down-stage and mid-stage counterparts. This comprehensive analysis through both quantitative metrics and qualitative visualizations demonstrates that upper-level features in the up-stage possess strong discriminative power for person search.

Layer	Detection		Re-ID	
	Recall	AP	mAP	Top-1
Down-stage Level0 Res0	95.4	91.1	42.1	83.2
Down-stage Level0 ViT0	95.9	91.7	43.3	84.1
Down-stage Level0 Res1	95.8	91.5	44.9	83.5
Down-stage Level0 ViT1	95.6	81.3	44.3	84.0
Down-stage Level0 Downsampler	95.1	91.1	42.4	82.5
Down-stage Level1 Res0	96.0	82.4	43.6	83.1
Down-stage Level1 ViT0	95.9	92.3	46.5	84.8
Down-stage Level1 Res1	96.2	92.7	47.3	84.9
Down-stage Level1 ViT1	**96.3**	**92.9**	**48.7**	**85.6**
Down-stage Level1 Downsampler	94.6	90.7	39.6	80.1
Down-stage Level2 Res0	95.0	91.3	42.7	81.2
Down-stage Level2 ViT0	94.9	91.4	43.5	82.5
Down-stage Level2 Res1	94.9	91.5	43.5	81.3
Down-stage Level2 ViT1	95.3	92.1	41.9	81.6
Down-stage Level2 Downsampler	91.1	83.2	9.6	43.3
Down-stage Level3 Res0	91.1	82.8	8.4	40.4
Down-stage Level3 Res1	90.0	81.7	6.9	36.4
Mid-stage Res0	90.2	81.5	6.4	33.7
Mid-stage ViT0	91	81.8	6.4	34
Mid-stage Res1	90.5	81.6	6.3	34.3

Table 9. Performance metrics for different layers in the down-stage and mid-stage of UNet on the PRW [23] dataset. We evaluate different feature maps obtained from Vision Transformer [4] (ViT) and ResNet [6] (Res) modules at each level. Each level contains multiple ViT and Res modules arranged sequentially, with the appended number (e.g., Res0, ViT0) indicating their order within that level. The downsampler represents feature maps from modules that reduce spatial resolution between adjacent levels. Numbers in bold indicate the best performance and underscored ones are the second best.

Timestep-wise analysis We show in Fig. 4 how feature representations evolve across different timesteps *(t)*. At *t=0*, features maintain clear semantic structure with precise person silhouettes, leading to optimal re-ID and detection performance. Features gradually degrade through intermediate timesteps (*t*=100-400), with person silhouettes becoming increasingly abstract. Later timesteps (*t*=500-1000) show severe degradation, with features becoming dominated by noise and losing meaningful patterns. Figure 4 shows this progression in detail. This analysis reveals that early timesteps (*t*=0-30) provide the most effective features for re-ID and detection tasks, informing our optimal timestep selection.

K. Limitation

In this work, we harness diffusion priors to person search and demonstrate their effectiveness. Our DiffPS leverages a pre-trained diffusion model as a large-scale foundation model, which could raise concerns about computational overhead. However, by adopting a frozen backbone, we maintain fewer learnable parameters compared to recent state-of-the-art models. Future research on efficient diffusion models could further address computational considerations while retaining our method's advantages.

Layer	Detection		Re-ID	
	Recall	AP	mAP	Top-1
Up-stage Level0 Res0	88.3	76.5	1.5	11.3
Up-stage Level0 Res1	89.5	78.3	1.6	12.2
Up-stage Level0 Res2	89.2	79.2	1.8	14.1
Up-stage Level0 Upsampler	88.7	79.4	1.1	8.9
Up-stage Level1 Res0	96.0	92.7	41.6	80.9
Up-stage Level1 ViT0 query	95.3	91.6	40.2	80.0
Up-stage Level1 ViT0 key	95.2	91.7	40.6	79.6
Up-stage Level1 ViT0 value	95.4	51.8	39.6	79.0
Up-stage Level1 ViT0	95.3	91.5	37.7	79.0
Up-stage Level1 Res1	96.1	92.7	44.9	82.6
Up-stage Level1 ViT1 query	95.8	92.6	42.3	81.2
Up-stage Level1 ViT1 key	95.9	92.4	41.5	80.5
Up-stage Level1 ViT1 value	95.7	92.3	42.7	81.7
Up-stage Level1 ViT1	95.9	92.6	40.7	80.6
Up-stage Level1 Res2	96.0	92.8	46.3	83.0
Up-stage Level1 ViT2 query	95.8	92.5	46.4	83.6
Up-stage Level1 ViT2 key	95.5	92.2	46.0	82.8
Up-stage Level1 ViT2 value	95.2	91.6	45.8	82.8
Up-stage Level1 ViT2	95.5	92.2	45.1	82.8
Up-stage Level1 Upsampler	96.8	93.7	39.4	80.5
Up-stage Level2 Res0	97.4	94.3	50.4	86.7
Up-stage Level2 ViT0 query	97.7	<u>94.7</u>	50.0	85.2
Up-stage Level2 ViT0 key	97.5	94.5	48.7	84.3
Up-stage Level2 ViT0 value	97.2	94.3	48.5	85.2
Up-stage Level2 ViT0	97.2	94.3	47.3	84.1
Up-stage Level2 Res1	97.6	94.6	53.7	86.4
Up-stage Level2 ViT1 query	97.6	94.5	53.6	85.9
Up-stage Level2 ViT1 key	97.6	94.5	52.3	85.7
Up-stage Level2 ViT1 value	97.5	94.1	<u>53.8</u>	87.1
Up-stage Level2 ViT1	97.3	94.2	52.1	86.5
Up-stage Level2 Res2	97.4	94.4	53.5	**87.9**
Up-stage Level2 ViT2 query	97.3	94.3	**54.4**	87.3
Up-stage Level2 ViT2 key	97.8	94.5	53.5	86.5
Up-stage Level2 ViT2 value	96.9	93.8	52.7	86.3
Up-stage Level2 ViT2	97.2	94.2	51.9	86.4
Up-stage Level2 Upsampler	97.8	<u>94.7</u>	51.9	86.5
Up-stage Level3 Res0	97.4	94.1	52.2	86.4
Up-stage Level3 ViT0 query	<u>98.0</u>	<u>94.7</u>	52.9	87.3
Up-stage Level3 ViT0 key	**98.1**	**94.8**	53.1	<u>87.7</u>
Up-stage Level3 ViT0 value	97.4	94.0	53.1	87.0
Up-stage Level3 ViT0	97.1	93.6	47.1	84.8
Up-stage Level3 Res1	97.7	94.3	52.0	86.2
Up-stage Level3 ViT1 query	97.5	94.1	52.8	86.6
Up-stage Level3 ViT1 key	97.5	94.2	53.2	86.8
Up-stage Level3 ViT1 value	97.5	94.0	52.7	86.3
Up-stage Level3 ViT1	97.4	93.8	48.1	85.0
Up-stage Level3 Res2	97.1	93.6	48.2	85.2
Up-stage Level3 ViT2 query	97.1	93.9	51.1	86.7
Up-stage Level3 ViT2 key	97.4	94.1	51.5	86.0
Up-stage Level3 ViT2 value	96.5	92.6	47.1	84.3
Up-stage Level3 ViT2	97.0	93.3	47.0	85.0

Table 10. Performance metrics for different layers in the up-stage of UNet on the PRW [23] dataset. We evaluate feature maps from Vision Transformer [4] (ViT) and ResNet [6] (Res) modules at each level. Each level contains multiple ViT and Res modules in sequence, with the appended number (*e.g.*, Res0, ViT0) indicating their order. For ViT modules, we analyze three attention-based feature maps (query, key, and value) after their linear projections. The upsampler represents feature maps from modules that increase spatial resolution between adjacent levels. Numbers in bold indicate the best performance and underscored ones are the second best.

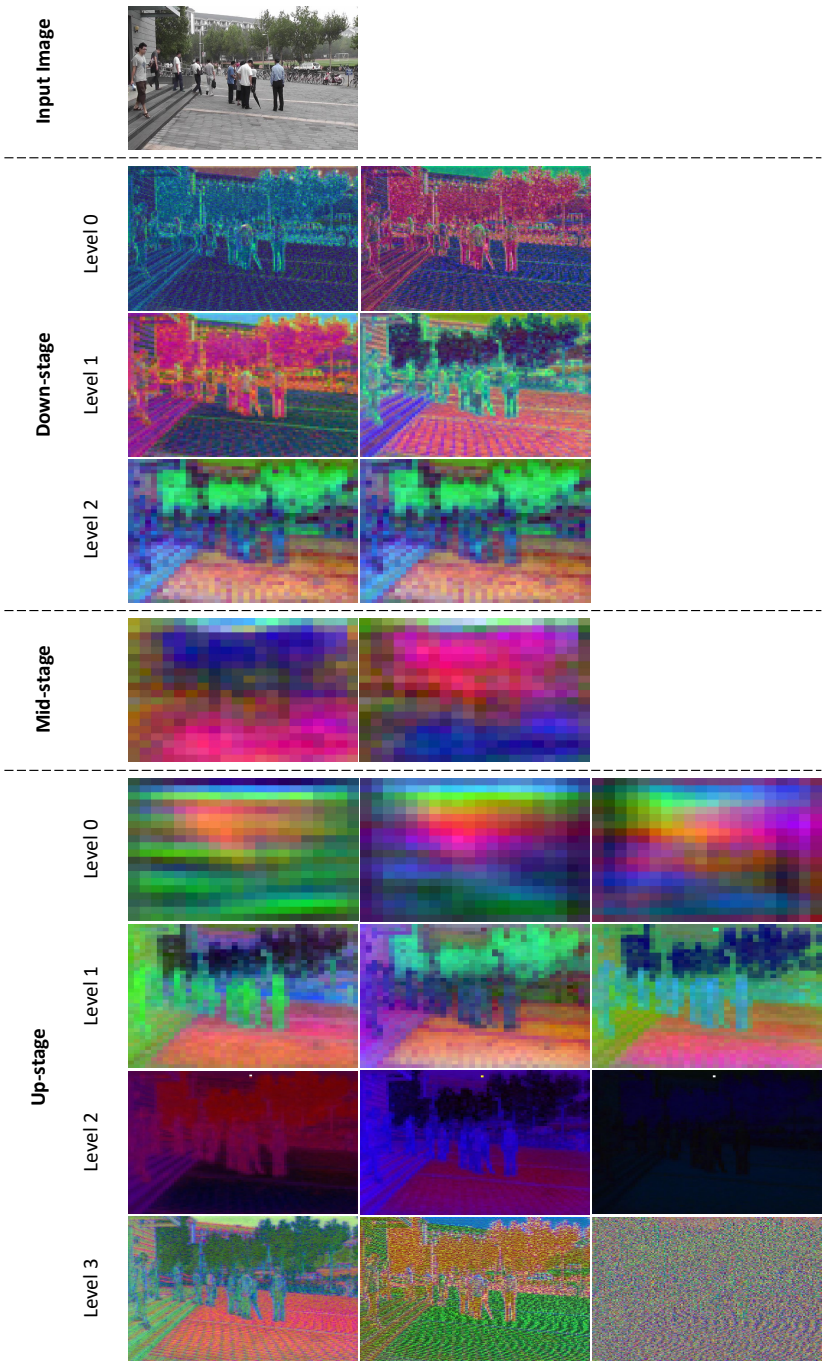


Figure 3. Feature map visualization from Res [6] modules across different stages and levels of UNet [17]. The visualizations are generated using PCA [12] on the output feature maps, with each row showing a different level and each column representing different res modules within that level. The input image is shown at the top for reference. Colors indicate the intensity and pattern of feature activations, demonstrating how feature representations evolve through different stages and levels of the network. (Best viewed in color.)

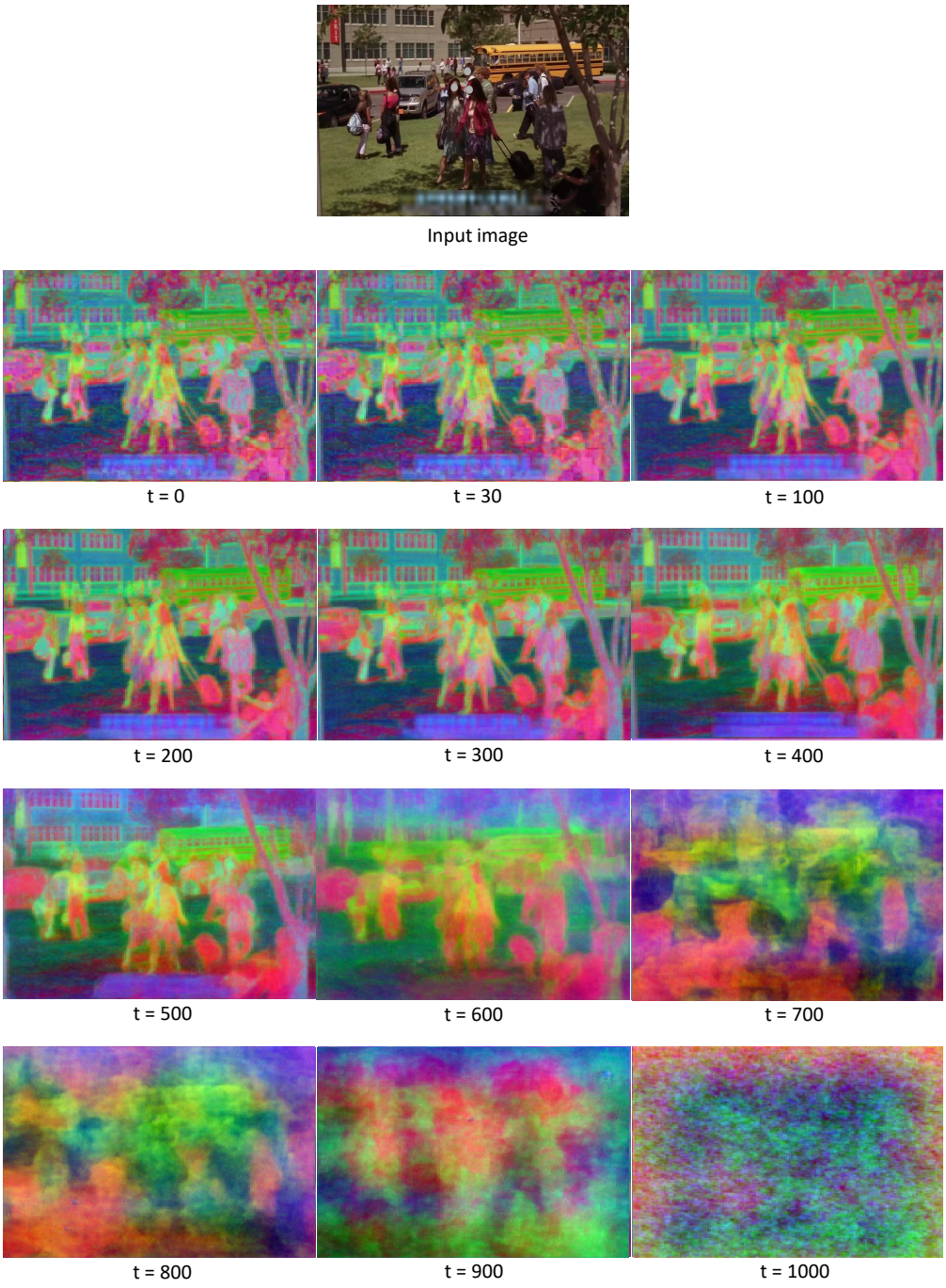


Figure 4. Visualization of feature characteristics across different timesteps in the diffusion process. Visualization using PCA [12] of features extracted from UNet [17] up-stage level 3 ViT [4] module at varying timesteps. The input image is shown at the top for reference. (Best viewed in color.)

References

[1] 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. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[3] 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

[4] 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. 1, 2, 4, 5, 7

[5] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2022. 3

[6] 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, 4, 5, 6

[7] 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. 2

[8] 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

[9] 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. 3

[10] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. 1

[11] 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

[12] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19 (3):303–342, 1993. 6, 7

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[14] 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

[15] 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. 1

[16] 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

[17] 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. 1, 4, 6, 7

[18] 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), 2018. 1

[19] 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

[20] Tianfu Wang, Guosheng Hu, and Hongguang Wang. Object pose estimation via the aggregation of diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10238–10247, 2024. 2

[21] 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. 2

[22] 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. 2, 3

[23] 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, 2, 3, 4, 5