GSAlign: Geometric and Semantic Alignment Network for Aerial-Ground Person Re-Identification

Qiao Li[1], Jie Li[2], Yukang Zhang[2], Lei Tan[3], Jing Chen[1], Jiayi Ji[2,3],

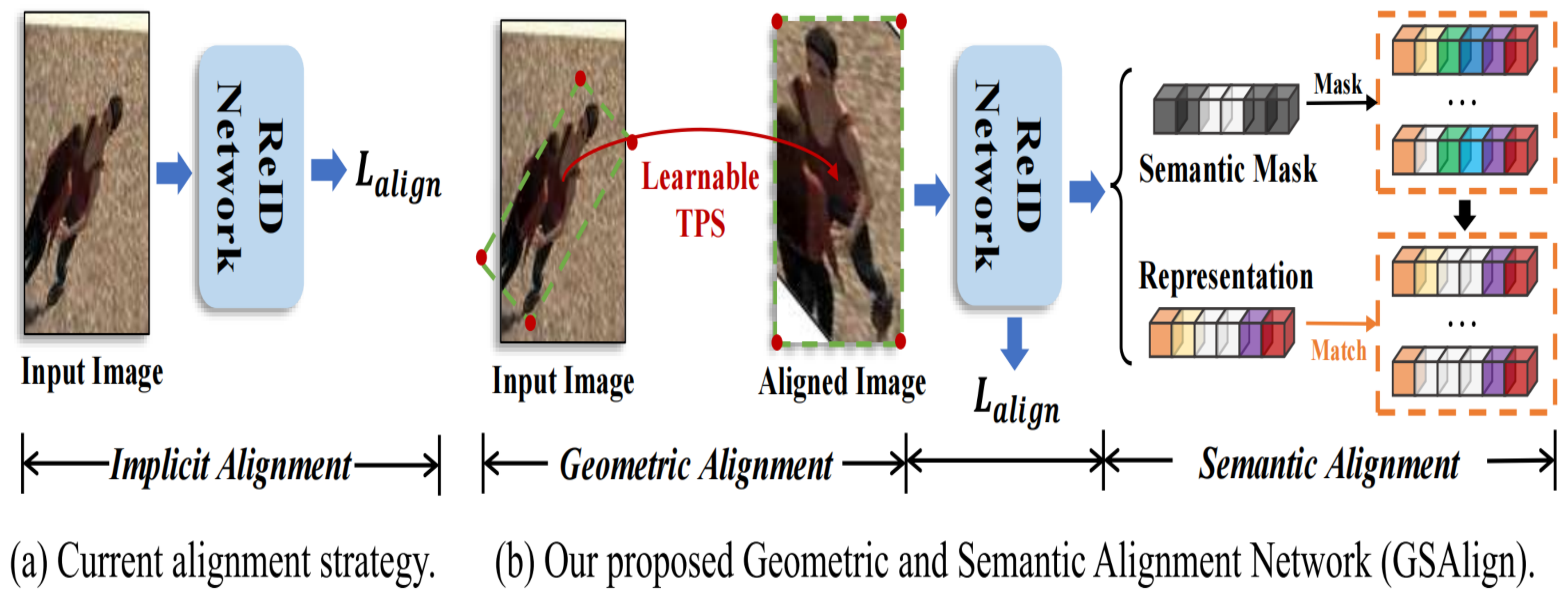
[1]School of Cyber Science and Engineering, Wuhan University, Wuhan, China

[2]School of Informatics, Xiamen University, Xiamen, China

[3]National University of Singapore, Singapore

NeurIPS | 2025

Motivation



(a) Current alignment strategy. (b) Our proposed Geometric and Semantic Alignment Network (GSAlign).

(a) Previous methods rely solely on implicit alignment, which is insufficient to fully address spatial and semantic distortions.

(b) In contrast, our GSAlign performs explicit alignment at both the geometric and semantic levels via LTPS and visibility-aware semantic masks, respectively. This design equips GSAlign with a stronger capability for robust aerial-ground matching.

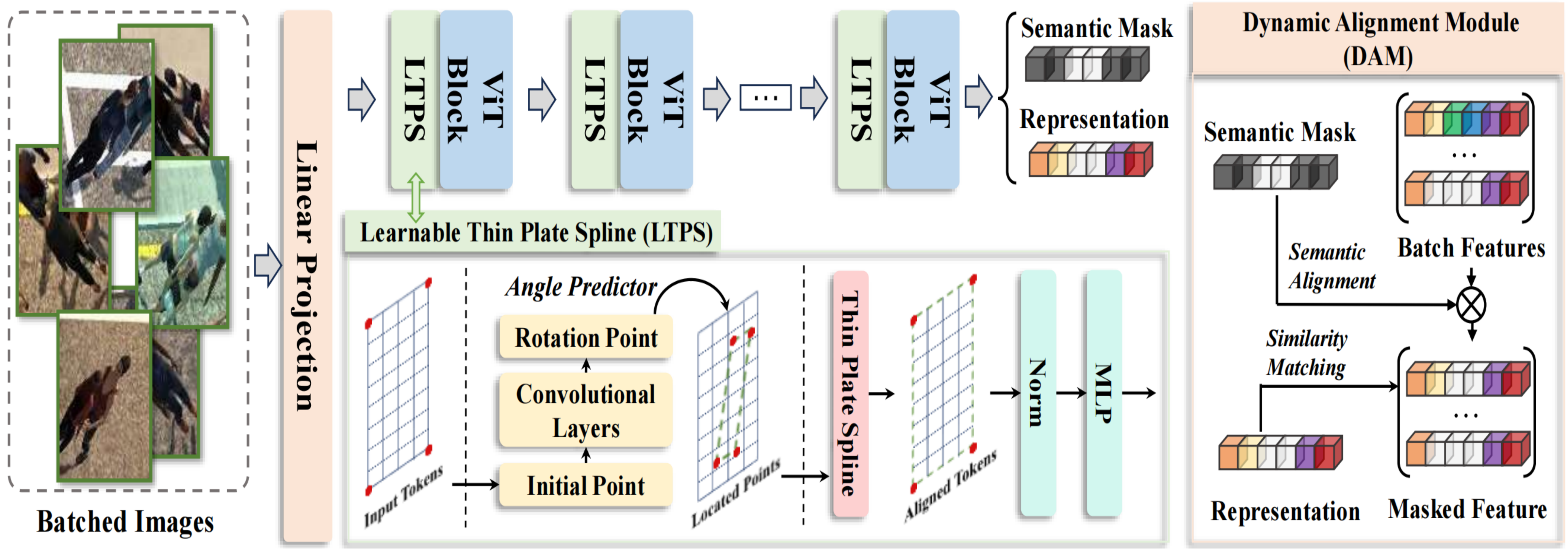
Contribution

(1) We propose GSAlign, a novel framework for aerial-ground person re-identification that jointly addresses geometric deformation and semantic misalignment within a unified architecture. GSAlign is specifically designed to handle the extreme cross-view variations and visibility inconsistencies inherent in UAV-to-ground matching scenarios.

(2) We introduce a Learnable Thin Plate Spline (LTPS) Module and a Dynamic Alignment Module (DAM). LTPS performs keypoint-guided feature warping to compensate for severe spatial distortions, while DAM enhances semantic alignment by estimating visibility-semantic representation masks to highlight visible body regions and suppress noisy or occluded areas.

(3) Extensive experiments on the challenging CARGO dataset validate the effectiveness of GSAlign, which achieves state-of-the-art performance with absolute gains of +18.8\% in mAP and +16.8\% in Rank-1 accuracy on the aerial-ground setting.

The Proposed GSAlign



GSAlign first applies an initial geometric transformation via a Learnable Thin Plate Spline (LTPS) module, followed by progressive alignment through LTPS blocks inserted before each ViT layer. In parallel, a Dynamic Alignment Module (DAM) generates a visibility-aware semantic mask according to the input image, which is then applied to the representations of other images in the batch to suppress irrelevant or occluded features.

Experiments

Method	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
SBS [40]	50.32	43.09	29.76	72.31	62.99	48.24	67.50	49.73	29.32	31.25	29.00	18.71
PCB [17]	51.00	44.50	32.20	74.10	67.60	55.10	55.00	44.60	27.00	34.40	30.40	20.10
BoT [41]	54.81	46.49	32.40	77.68	66.47	51.34	65.00	49.79	29.82	36.25	32.56	21.46
MGN [42]	54.81	49.08	36.52	**83.93**	71.05	55.20	65.00	52.96	36.78	31.87	33.47	24.64
VV [43, 44]	45.83	38.84	39.57	72.31	62.99	48.24	67.50	49.73	29.32	31.25	29.00	18.71
AGW [39]	60.26	53.44	40.22	81.25	71.66	58.09	67.50	56.48	40.40	43.57	40.90	29.39
BAU [45]	45.20	38.40	-	61.60	51.20	-	50.00	42.60	-	40.40	36.70	-
PAT [46]	37.90	15.30	-	52.70	24.20	-	50.00	23.10	-	35.10	15.50	-
DTST [47]	64.42	55.73	41.92	78.57	72.40	62.10	80.00	63.31	44.67	50.53	43.49	29.46
ViT [37]	61.54	53.54	39.62	82.14	71.34	57.55	80.00	64.47	47.07	43.13	40.11	28.20
VDT [8]	64.10	55.20	41.13	82.14	71.59	58.39	**82.50**	**66.83**	**50.22**	48.12	42.76	29.95
GSAlign	**65.06**	**57.95**	**44.97**	83.04	**73.86**	**62.73**	80.00	65.55	49.81	**64.89**	**61.55**	**52.81**

Setting	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
Baseline	64.10	55.20	41.13	82.14	71.59	58.39	**82.50**	**66.83**	**50.22**	48.12	42.76	29.95
Baseline + LTPS	64.42	55.95	41.92	80.36	71.87	59.55	**82.50**	65.26	47.15	**64.89**	61.08	50.54
Baseline + LTPS + DAM	**65.06**	**57.95**	**44.97**	**83.04**	**73.86**	**62.73**	80.00	65.55	49.81	**64.89**	**61.55**	**52.81**

Setting	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
Different variants of DAM												
Inner-Batch	65.06	**57.95**	**44.97**	**83.04**	73.86	**62.73**	**80.00**	**65.55**	**49.81**	**64.89**	**61.55**	**52.81**
Memory Bank	**65.38**	57.34	44.09	**83.04**	73.72	62.05	**80.00**	62.70	43.88	63.83	61.06	52.52
Classification Matrix	63.14	55.64	42.07	81.25	72.04	59.53	75.00	63.79	48.06	57.45	56.55	47.33

Setting	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
Different locations for LTPS												
First layer	64.10	55.92	42.44	**83.04**	72.86	60.58	**80.00**	**65.98**	**50.45**	58.51	56.92	47.62
First 4 layers	64.10	56.46	43.50	81.25	74.49	64.70	**80.00**	64.45	47.11	58.51	56.21	46.66
Middle 4 layers	64.74	57.09	44.08	82.14	**74.93**	**64.77**	77.50	64.28	47.30	58.51	58.30	50.18
Last 4 layers	**65.06**	57.39	44.05	**83.04**	74.42	62.86	77.50	65.21	49.80	**64.89**	59.87	50.95
All layers	**65.06**	**57.95**	**44.97**	**83.04**	73.86	62.73	**80.00**	65.55	49.81	**64.89**	**61.55**	**52.81**

Setting	Protocol 1: A↔G		Protocol 2: G↔A		Protocol 3: A↔W		Protocol 4: W↔A	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
BoT [41]	85.40	77.03	84.65	75.90	89.77	80.48	84.65	76.90
Explain [6]	87.70	79.00	87.35	78.24	93.67	83.14	87.73	79.08
VDT [8]	86.46	79.13	86.14	78.12	90.00	82.21	85.26	78.52
AG-ReIDv2 [16]	88.77	80.72	87.86	78.51	**93.62**	84.85	**88.61**	80.11
SeCap [49]	88.12	80.84	88.24	79.99	91.44	84.01	87.56	80.15
GSAlign	**91.47**	**89.78**	**88.29**	**87.62**	93.30	**91.84**	88.12	**88.62**

Visulization

The input image (red) exhibits significant geometric distortion due to extreme viewpoint variation. After applying the Learnable Thin Plate Spline (LTPS) transformation (green), the image is spatially rectified, highlighting improved geometric consistency and local structure alignment