Uncertainty-Aware Alignment Network for Cross-Domain Video-Text Retrieval

Xiaoshuai Hao[1], Wanqian Zhang[2]

[1]SAIT-China Lab, Samsung Research Center [2] Institute of Information Engineering, Chinese Academy of Sciences

xshuai.hao@samsunng.com, zhangwanqian@iie.ac.cn

Section 1: Motivation/Contribution

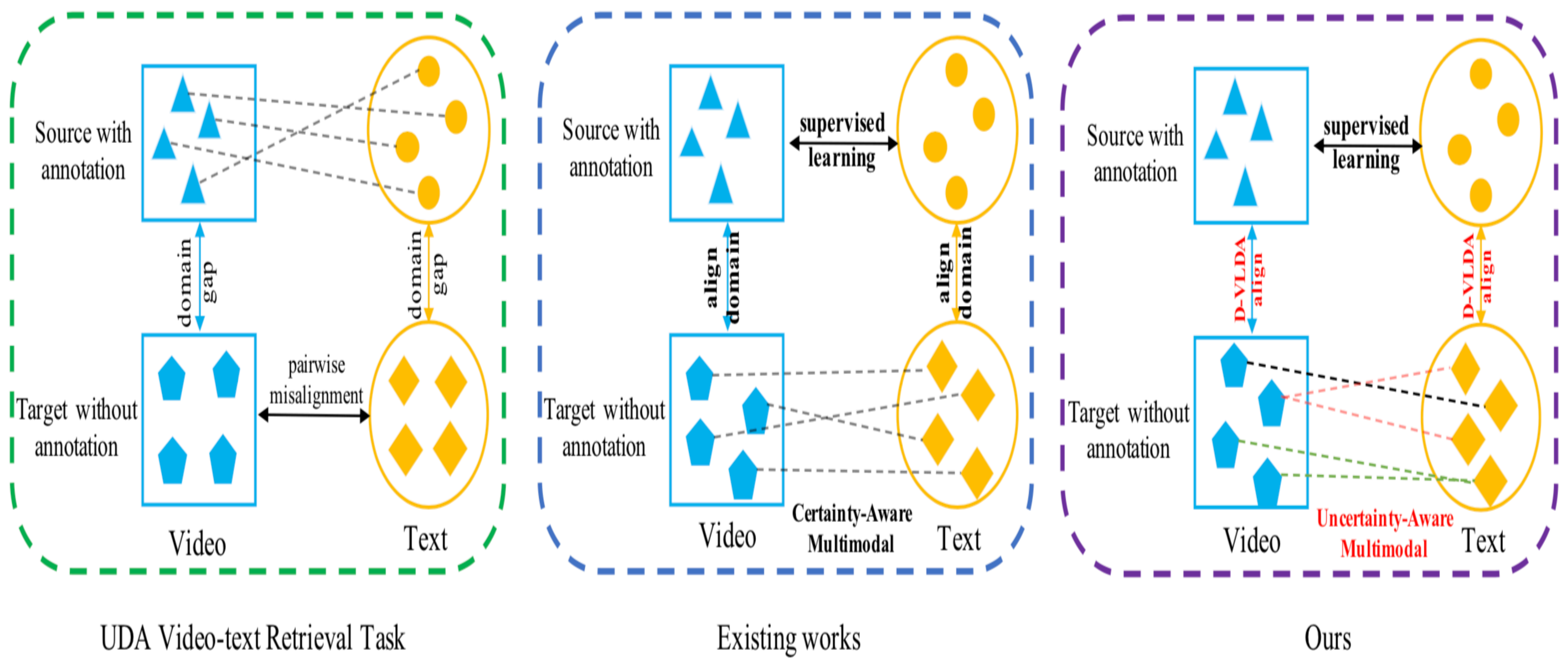


Figure 1: Illustration of the UDA Video-text Retrieval Task, existing works and the proposed method. The proposed method uses Distribution-based Vision-Language Domain Adaptation(D-VLDA) in domain gap and uncertainty-aware multimodal alignment mechanism in the target domain.

Motivation:

- Video-text retrieval is an important but challenging research task in the multimedia community. In this paper, we address the challenge task of Unsupervised Domain Adaptation Video-text Retrieval (UDAVR), assuming that training (source) data and testing (target) data are from different domains.
- Previous approaches are mostly derived from classification-based domain adaptation methods, which are neither multi-modal nor suitable for retrieval task. In addition, as to the pairwise misalignment issue in target domain, i.e., no pair annotations between target videos and texts, the existing method assumes that a video corresponds to a text. Yet we empirically find that in the real scene, one text usually corresponds to multiple videos and vice versa.
- To tackle this one-to-many issue, we propose a novel method named Uncertainty-aware Alignment Network (UAN). Specifically, we first introduce the multi-modal mutual information module to balance the minimization of domain shift in a smooth manner. To tackle the multimodal uncertainties pairwise misalignment in target domain, we propose the Uncertainty-aware Alignment Mechanism (UAM) to fully exploit the semantic information of both modalities in target domain.

Contribution:

- For the challenging Unsupervised Domain Adaptation Video-text Retrieval (UDAVR) task, we propose a simple yet effective Uncertainty-aware Alignment Network (UAN), which fully exploits the semantic information of both modalities in target domain.
- To tackle the one-to-many in target domain, the proposed Uncertainty-aware Alignment Mechanism (UAM) tries to utilize the multi-granularity relationships between each target video and text to ensures the discriminability of target features.
- Compared with the state-of-the-art methods, UAN achieves 15.47% and 11.57% relative improvements on R@1 under the setting of TGIF→MSRVTT and MSRVTT→TGIF respectively, demonstrating the superiority of our method.

Section 2: Method

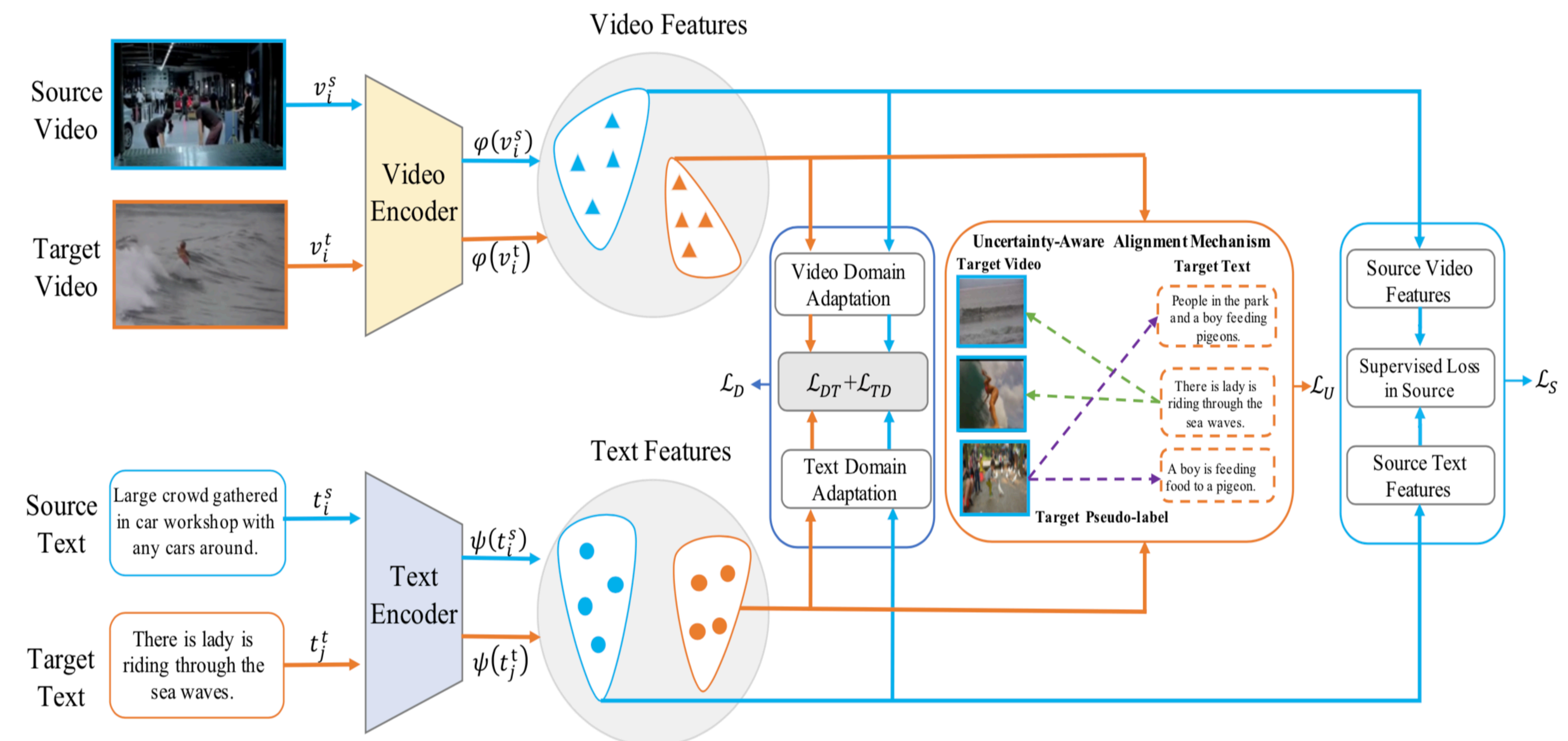


Figure 2: The overall framework of UAN. Semantic Embedding Learning to generate discriminative source features in a joint embedding space($\mathcal{L}_S$). Distribution-based Vision-Language Domain Adaptation(D-VLDA) is proposed to alleviate the domain discrepancy problem in both modalities($\mathcal{L}_D$). Uncertainty-Aware Alignment Mechanism(UAM) is proposed to dig in uncertainty-aware multimodal relationships in the target domain($\mathcal{L}_U$).

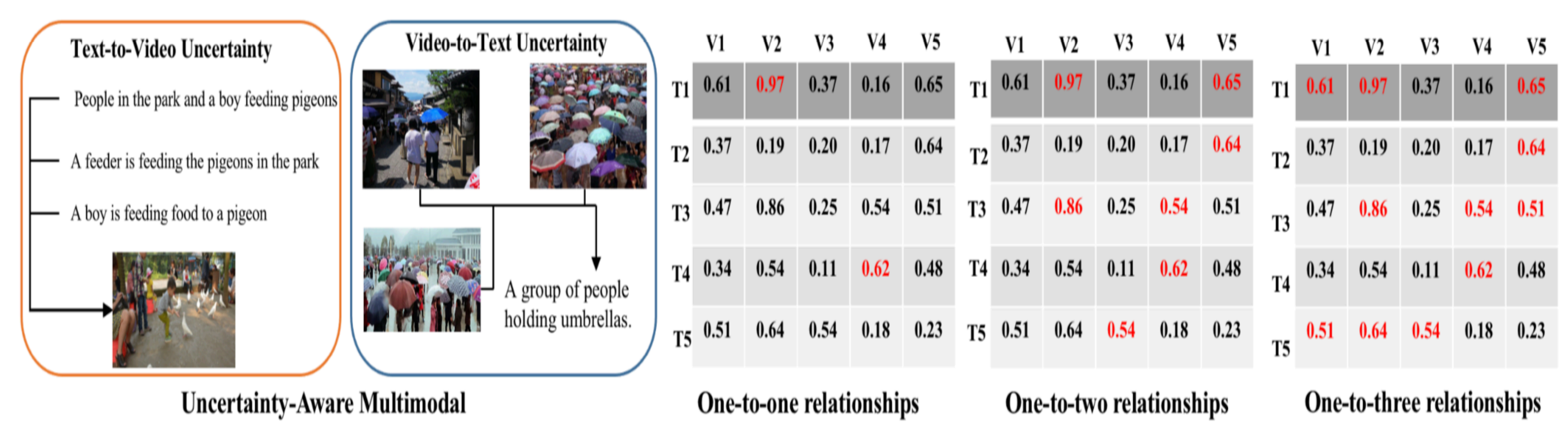
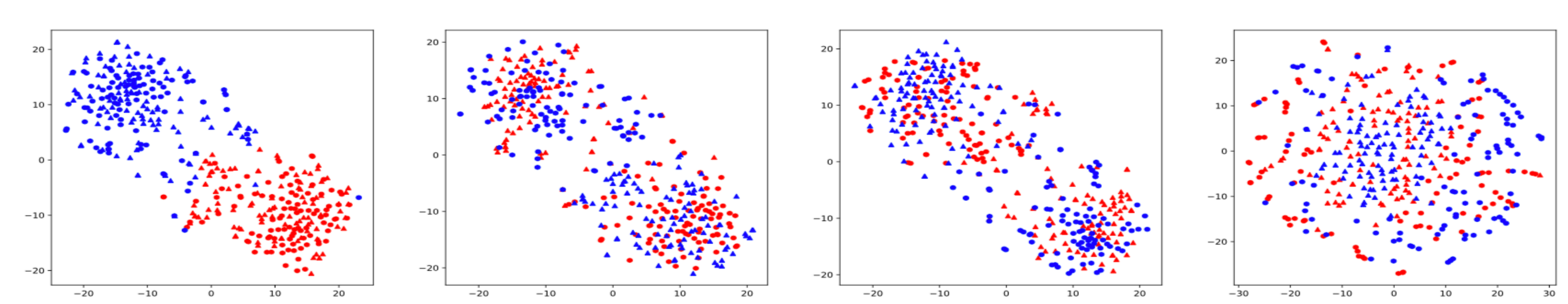


Figure 3: Illustration of Uncertainty-Aware Alignment Mechanism (UAM). If v^t_{i*} and t^t_{j*} to be the reciprocal *TOP-K* similar of each other, indicating a truly aligned (or positive) pair.



(a) Source Only (b) MMD [38] (c) ACP [37] (d) Our method

Figure 5: The t-SNE visualizations of (a) Source Only, (b) MMD, (c) ACP and (d) Our method.

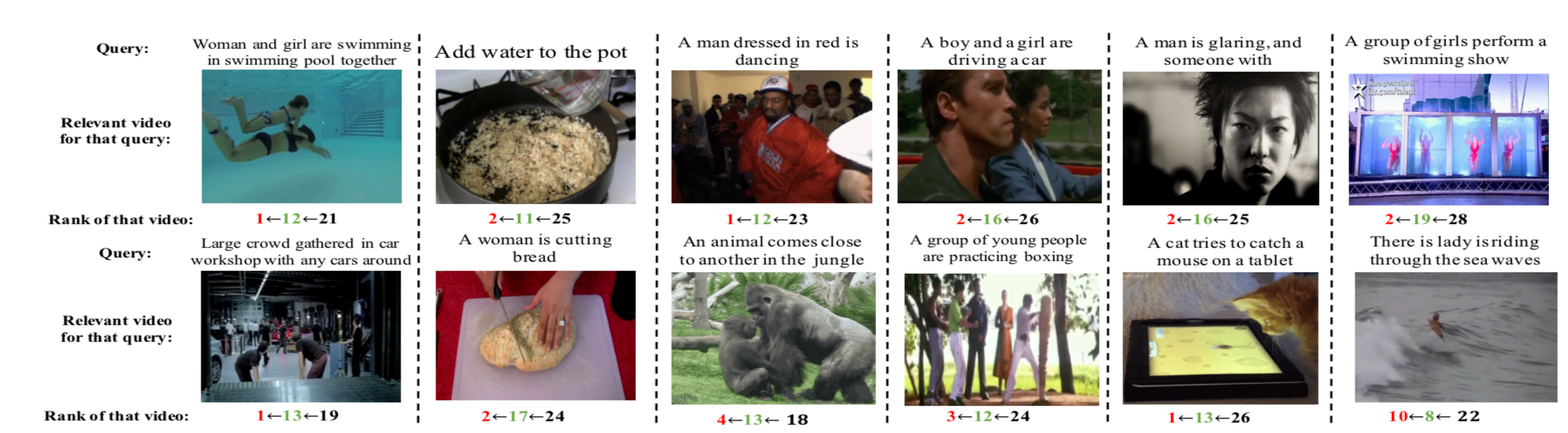


Figure 6: Visualizations of video-text results. Qualitative results of query texts and corresponding videos along with the changes in rank A←B←C, where A denotes the rank of UAN, B the ACP method and C the Source Only.

Section 3: Experiments

Table 1: Comparison with different baselines.

Method		Tf→Mt			Mt→Tf			Tf→Md			Md→Tf			Mt→Md			Md→Mt		
		R1↑	R10↑	MR↓	R1↑	R10↑	MR↓	R1↑	R10↑	MR↓	R1↑	R10↑	MR↓	R1↑	R10↑	MR↓	R1↑	R10↑	MR↓
Source Only		2.69	13.63	144	6.30	25.43	60	9.39	37.77	20	3.80	16.99	102	15.02	46.96	12	2.50	13.27	136
(a)	MMD [38]	2.68	13.59	135	6.77	27.11	54	9.11	36.11	23	3.50	16.28	119	15.31	47.65	12	2.62	13.18	136
	CORAL [49]	2.74	14.07	128	6.56	26.49	52	9.44	37.87	21	3.65	17.34	108	15.65	49.43	11	2.65	13.34	138
	DANN [16]	2.76	13.94	127	6.86	27.17	48	9.27	38.00	20	3.74	16.72	103	15.67	48.67	11	2.62	13.17	134
	IDM [11]	2.59	13.11	149	7.12	25.35	60	8.05	35.51	23	3.24	15.78	120	13.96	47.77	12	2.54	12.39	165
	SCDA [32]	2.79	14.22	130	6.92	26.70	53	9.84	37.11	22	3.30	17.02	108	15.64	48.65	11	2.55	12.98	138
(b)	MAN [13]	2.53	12.98	144	6.42	25.96	63	8.84	37.06	21	3.06	16.31	119	15.05	48.51	11	2.40	12.00	174
	CAPQ [6]	3.46	17.02	110	7.33	25.64	62	9.30	37.97	21	3.97	17.75	113	15.66	49.08	11	3.35	15.47	158
	ACP [37]	4.41	21.72	64	7.83	26.72	50	12.09	41.38	18	5.12	21.46	82	17.87	54.34	8	5.90	25.68	54
	DADA [22]	5.30	24.54	50	8.21	28.97	45	14.34	48.77	11	6.03	22.52	78	18.97	57.93	7	6.40	27.61	42
Ours	**UAN**	**6.12**	**27.23**	**40**	**9.16**	**31.06**	**37**	**15.15**	**49.34**	**10**	**6.51**	**23.93**	**69**	**20.25**	**60.23**	**5**	**6.52**	**28.15**	**40**

Table 2: Effect of $\mathcal{L}_D$ and $\mathcal{L}_U$.

Method	Tf→Mt			Mt→Tf		
	R1↑	R10↑	MR↓	R1↑	R10↑	MR↓
Source Only	2.69	13.63	144	6.30	25.43	60
UAN(w/o $\mathcal{L}_D$)	3.63	18.12	79	6.72	26.76	50
UAN(w/o $\mathcal{L}_U$)	3.87	18.45	76	6.91	27.01	48
UAN(full)	**6.12**	**27.23**	**40**	**9.16**	**31.06**	**37**

Table 3: Analysis on different alignment mechanisms.

Method	Tf→Mt			Mt→Tf		
	R1↑	R10↑	MR↓	R1↑	R10↑	MR↓
UAN(w/ DAC)	5.43	25.63	48	8.44	29.12	42
UAN(w/ UAM)	5.76	26.16	43	8.73	29.84	40
UAN(full)	**6.12**	**27.23**	**40**	**9.16**	**31.06**	**37**

Table 5: Generalization to different video-text retrieval methods.

Method		Tf→Mt			Mt→Tf		
		R1↑	R10↑	MR↓	R1↑	R10↑	MR↓
(a)	HGR [7]	2.20	11.98	154	5.87	22.10	72
	HGR + UAN	**4.52**	**21.32**	**86**	**8.43**	**29.23**	**45**
	GPO [5]	2.69	13.63	144	6.30	25.43	60
	GPO + UAN	**6.12**	**27.23**	**40**	**9.16**	**31.06**	**37**
(b)	CE [35]	2.93	14.7	122	6.50	26.23	56
	CE + UAN	**6.23**	**27.25**	**41**	**9.32**	**32.41**	**34**
	MMT [15]	4.20	22.30	78	7.32	31.46	30
	MMT + UAN	**6.53**	**28.32**	**34**	**9.63**	**38.89**	**19**
(c)	CLIP4CLIP [42]	7.20	28.50	35	10.43	38.16	26
	CLIP4CLIP + UAN	**9.32**	**37.85**	**22**	**13.55**	**47.33**	**15**
	CLIP2Video [14]	7.80	31.50	31	11.21	39.48	23
	CLIP2Video + UAN	**9.72**	**38.43**	**17**	**14.23**	**47.87**	**12**

Table 4: Analysis on different DA methods.

Method	Tf→Mt			Mt→Tf		
	R1↑	R10↑	MR↓	R1↑	R10↑	MR↓
Source Only	2.69	13.63	144	6.30	25.43	60
UAN(w/ MMD [38])	5.62	25.56	45	8.36	29.61	41
UAN(w/ CORAL [49])	5.64	25.58	44	8.38	29.64	41
UAN(w/ GRL [17])	5.66	25.67	44	8.43	29.71	40
UAN(w/ TPN [46])	5.72	25.72	43	8.46	29.75	40
UAN(w/ CDD [28])	5.75	25.73	43	8.56	29.82	39
UAN(w/ MSTN [58])	5.76	25.82	42	8.58	29.91	39
UAN(w/ D-VLDA)	**6.12**	**27.23**	**40**	**9.16**	**31.06**	**37**

Table 6: The results of image-text retrieval.

Method		Open Narr→ COCO Narr			COCO→ COCO Narr		
		R1↑	R10↑	MR↓	R1↑	R10↑	MR↓
(a)	SCAN [29]	17.4	52.6	9	22.3	72.98	5
	VSRN [31]	19.6	54.7	7	25.1	75.4	4
	CE [35]	19.6	56.4	7	24.5	75.8	4
(b)	CDAN [39]	20.6	59.2	6	22.2	73.3	5
	CORAL [49]	19.4	58.3	7	25.4	74.6	4
	DANN [16]	19.0	58.4	7	24.8	76.8	4
	MMD [38]	17.3	50.8	9	22.6	72.0	5
	OT [57]	20.3	57.1	8	25.0	75.6	4
(c)	MAN [13]	20.4	57.3	8	25.6	75.8	4
	CAPQ [6]	21.8	57.4	7	26.5	76.4	4
	ACP [37]	22.3	57.9	6	27.3	77.9	4
	DADA [22]	22.9	58.3	5	28.1	78.3	4
Ours	**UAN**	**23.6**	**59.2**	**4**	**29.5**	**79.5**	**3**

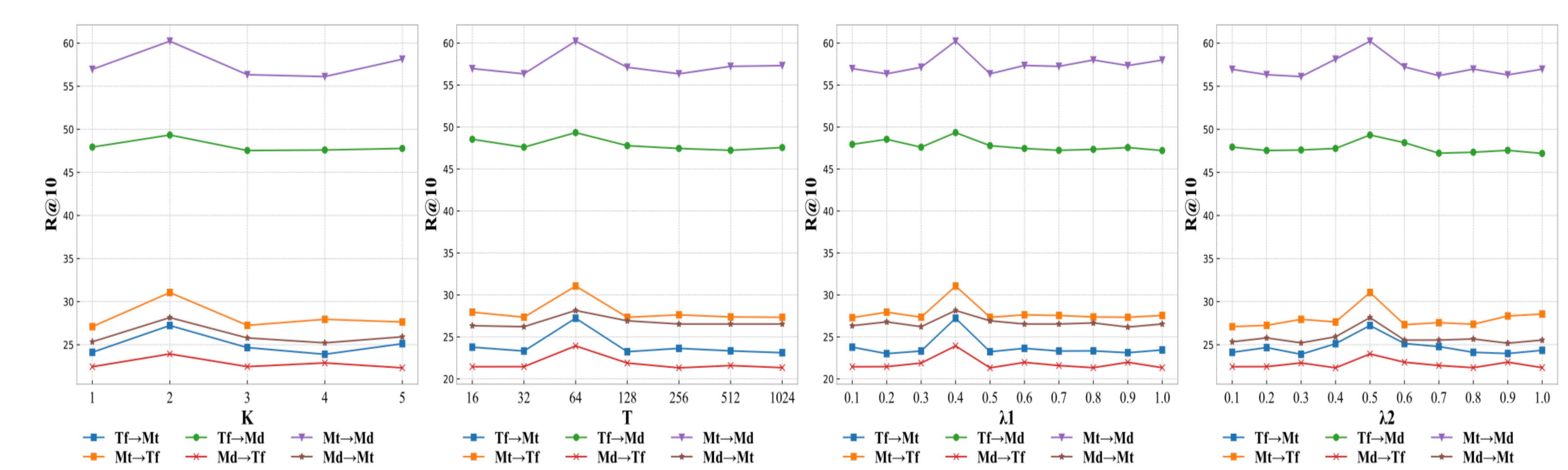


Figure 4: Analysis of hyper parameters, i.e., K, T, λ_1, and λ_2, with the R@10 retrieval performance across different domain adaptation directions.