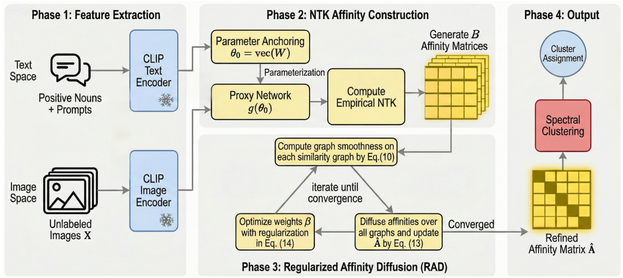
DELVING INTO SPECTRAL CLUSTERING WITH VISIONLANGUAGE REPRESENTATIONS

Bo Peng[1], Yuanwei Hu[1], Bo Liu[1], Ling Chen[1], Jie Lu[1], Zhen Fang[1]

1 Australian Artificial Intelligence Institute, University of Technology Sydney, Australia

Why This Matters

We reveal **spectral clustering's blind spot**: relying on single-modal affinities blurs semantics when images look alike. Leveraging vision–language signals, our approach **injects text-informed structure** into affinities, sharpening clusters with **clearer blocks** and lower noise.



Core Idea In One Line

We propose **Neural Tangent Kernel Spectral Clustering** that couples visual proximity and semantic overlap. Anchored by **positive nouns** from the wild, our NTK transforms CLIP features into **block-diagonal** affinities that drive stronger, cleaner partitions.

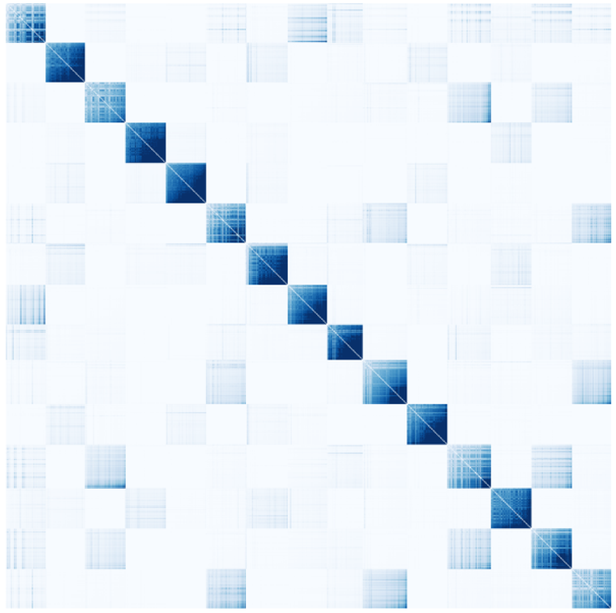
NTK Construction

- Extract CLIP text embeddings W from filtered **positive nouns**. Anchor NTK with *vec(W)* to align gradients with **text semantics**. Use log-sum-exp proxy: gradients capture how images align to anchors, yielding **text-informed** affinities.

Dataset	STL-10			CIFAR-10			CIFAR-20			ImageNet-10			ImageNet-Dogs			DTD			UCF101			ImageNet-1K			Average		
Metrics	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**
TAC (Kmeans)	92.3	94.5	89.5	80.8	90.1	79.8	60.7	55.8	42.7	97.5	98.6	97.0	75.1	75.1	63.6	60.1	45.9	29.0	81.6	61.3	52.4	77.8	48.9	36.4	78.3	71.3	61.3
TAC (SC)	92.6	94.3	94.2	81.2	90.3	80.1	56.9	54.5	30.1	97.0	98.3	96.8	75.3	75.8	64.4	58.6	44.0	27.1	79.6	60.0	50.1	78.0	49.1	36.2	77.4	70.8	59.9
Ours (naive)	87.0	91.2	84.0	71.4	74.7	56.8	45.1	42.6	29.9	86.1	90.4	80.9	70.5	69.0	56.1	56.2	45.6	30.2	82.2	64.9	57.9	77.1	51.0	35.4	72.0	66.2	53.9
Ours (PE)	93.1	97.9	89.6	82.9	91.3	82.4	60.9	55.0	43.4	97.6	98.9	97.4	82.3	81.3	72.3	60.7	50.6	32.1	81.5	66.8	58.9	79.2	53.3	38.4	79.8	74.4	64.3
Ours (RAD)	**95.8**	**98.3**	**96.3**	**83.3**	**92.0**	**83.0**	**63.3**	**59.6**	**43.5**	**97.8**	**99.2**	**98.4**	**82.4**	**84.9**	**71.4**	**61.7**	**52.0**	**33.6**	**83.0**	**67.9**	**59.4**	**79.2**	**56.3**	**39.4**	**80.8**	**76.3**	**65.6**

Coupling That Separates

Affinity equals *visual proximity U_ij* times *semantic overlap V_ij*. Within-cluster pairs share **high** *U_ij* and *V_ij*, boosting links; cross-cluster pairs lack shared nouns, so **low** *V_ij* suppresses spurious edges. Result: **crisp block-diagonal** A_NTK.



(c) Ours (NMI=82.4%)

Multi-Prompt Power: RAD

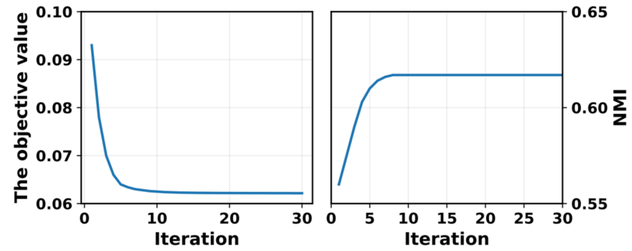
We introduce **Regularized Affinity Diffusion (RAD)** to ensemble prompt-specific affinities. It jointly optimizes weights β and diffusion, avoids over-smoothing, and converges fast. Outcome: a **robust** affinity ^A that exploits **complementary** prompts.

Results That Stand Out

Our framework boosts clustering: 98.3% ACC on STL-10, 84.9% ACC on ImageNet-Dogs; gains over TAC by +3.8%/+9.8%. On DTD, UCF-101, ImageNet-1K, we average +7.7% NMI, +2.5% ACC, +6.3% ARI. Fine-grained: **+5.1%** (Pets ACC). Domain-shift: **higher** robustness.

Sharper Graphs, Faster Convergence

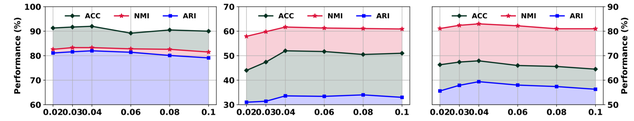
We reveal **sharply block-diagonal** A_NTK vs. RBF on CLIP or TAC features. RAD's objective **monotonically drops**; NMI **rises** and stabilizes in few iterations, delivering efficient optimization and **reliable** clustering.



(b) DTD

Ablations & Robustness

- Stable across τ, q, μ, λ; larger backbones further lift scores. Single-prompt still **beats** TAC; gains not from prompt engineering. Domain variants (ImageNet-C/V2/S) show **consistent** improvements; shifts remain challenging.



Takeaways & Limits

Our method **unlocks multi-modal spectral clustering**: NTK anchored by text **amplifies semantics**, RAD **adapts prompts**. Limitation: requires known cluster count. Future: **automatic K**, richer anchors, scalable diffusion—toward robust, text-aware clustering.

Backbone	Dataset	STL-10			CIFAR-10			CIFAR-20			ImageNet-Dogs			DTD		
	Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
ViT-B/16	TAC (KMeans)	95.1	96.1	93.6	82.9	91.2	80.5	62.6	60.4	45.7	82.3	81.9	72.9	62.6	50.4	33.6
	TAC (SC)	92.9	96.3	92.3	83.1	91.3	82.0	63.0	60.0	45.1	82.1	81.5	73.0	63.1	51.7	34.6
	Ours	**97.2**	**99.0**	**97.4**	**86.0**	**93.1**	**85.4**	**65.7**	**64.4**	**49.1**	**84.9**	**86.7**	**75.0**	**66.3**	**55.8**	**37.2**
ViT-L/14	TAC (KMeans)	95.4	96.7	94.2	89.1	93.9	86.7	64.8	62.9	47.6	84.3	84.0	75.4	64.7	52.9	35.1
	TAC (SC)	93.8	96.7	93.8	89.4	94.1	88.0	65.0	63.0	47.5	84.0	84.0	75.0	65.1	53.5	35.9
	Ours	**97.7**	**99.5**	**98.1**	**92.1**	**96.6**	**90.7**	**66.9**	**66.1**	**50.8**	**86.2**	**88.3**	**79.0**	**68.1**	**58.0**	**38.9**