FANS: A Flatness-Aware Network Structure for Generalization in Offline RL

Da Wang[1], Yi Ma[1]*, Ting Guo[2], Hongyao Tang[3], Wei Wei[1], Jiye Liang[1]

[1] School of Computer and Information Technology, Shanxi University. [2] Data Science and Technology, North University of China.
[3] College of Intelligence and Computing, Tianjin University

Background：Offline Reinforcement Learning

Sergey Levine，NeurIPS 2020 Tutorial

1. **Policy constraint**
2. **Model-based**
3. **Value-regularized**
4. **Uncertainty-based**

Method	Modifies	Modification type	Extra requirement
Policy constraints	Learned policy	Distribution, support, state-marginal	Behavior policy estimate
Model-based	Reward function	Uncertainty penalty	Uncertainty/ conservatism metric
Value regularization	Q-function	Q-function objective	Regularizer
Uncertainty	Q-function, model	Weighting update	Uncertainty estimate

Conservative → Avoid being too conservative

CQL (NeurIPS, 2020)
TD3BC (NeurIPS, 2021)
……

MCQ (NeurIPS, 2022)
POR (NeurIPS, 2022)
PRDC (ICML, 2023)
STR (ICML, 2023)
ADS (ICML, 2024)
LAD (AAAI, 2025)

Existing methods often relax constraints on policy and data or extract informative patterns through data-driven techniques. However, there has been limited exploration into **structurally guiding the optimization process toward flatter regions** of the solution space that **offer better generalization**.

Contribution

1. We propose a structured network design framework for offline RL that integrates residual blocks, Gaussian activation function, layer normalization, and ensemble techniques to enhance generalization.
2. We validate the effectiveness of the proposed framework across multiple offline RL tasks, highlighting that our remarkably simple architecture leads to substantial performance gains.
3. We conduct a detailed analysis to elucidate how the FANS framework facilitates smoother optimization, reduces variance, and mitigates overfitting, thereby achieving significant improvements in OOD generalization performance.

FANS framework: Flatness-Aware Network Structure

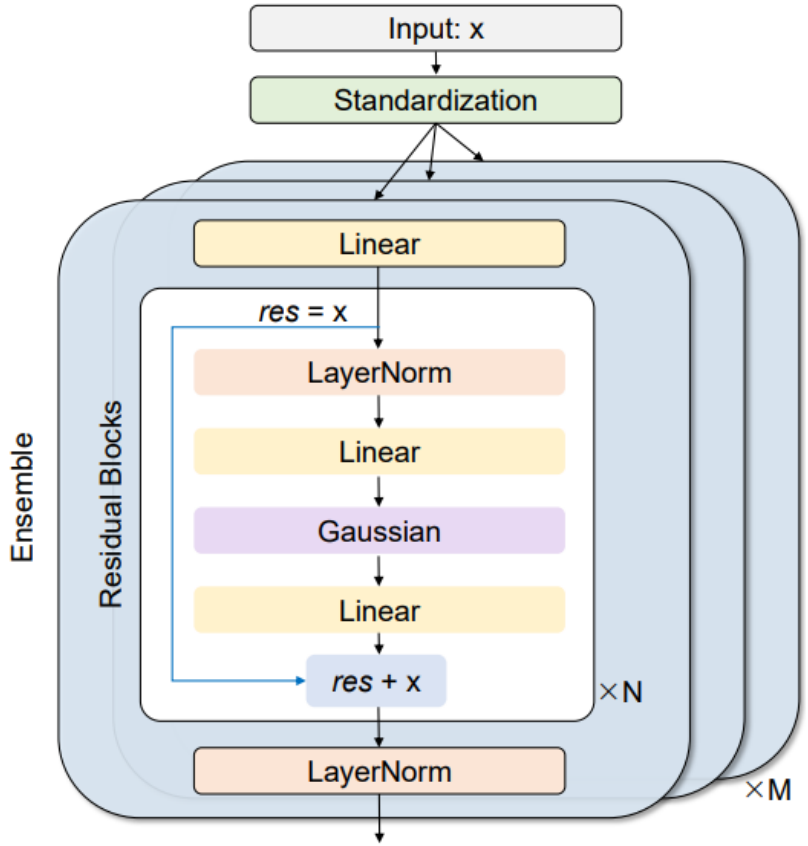


Figure 1: **FANS architecture.**

Residual Block

Table 1: Residual Block Architecture in FANS. Each step operates on a hidden vector of dimension d.

Step	Operation	Equation	Description
(1)	Residual Save	$\mathbf{res} = \mathbf{x}$	Store input for residual connection
(2)	LayerNorm	$\mathbf{h}_1 = \text{LayerNorm}(\mathbf{x})$	Normalize input across features
(3)	Linear Layer 1	$\mathbf{h}_2 = \mathbf{W}_1\mathbf{h}_1 + \mathbf{b}_1$	First linear transformation
(4)	Gaussian Activation	$\mathbf{h}_3 = \exp(-\mathbf{h}_2^2)$	Smooth, non-monotonic nonlinearity
(5)	Linear Layer 2	$\mathbf{h}_4 = \mathbf{W}_2\mathbf{h}_3 + \mathbf{b}_2$	Second linear transformation
(6)	Residual Add	$\mathbf{y} = \mathbf{res} + \mathbf{h}_4$	Residual connection output

Gaussian activation

$$\phi(u) = \exp(-u^2)$$

Layer normalization

$$\mathbf{z} = \text{LayerNorm}(\mathbf{y})$$

Ensemble

$$Q_{\text{ensemble}}(s,a) = \frac{1}{M}\sum_{m=1}^{M} Q^{(m)}(s,a)$$

Experiments on the D4RL Benchmark

Main Results

Table 2: Performance comparison on D4RL locomotion tasks over the final ten evaluations and five seeds (normalized scores). We **bold** the highest mean.

Tasks	TD3BC	AWAC	CQL	IQL	ReBRAC	SAC-N	EDAC	DT	**TD3 +FANS**
ha-m	48.1 ±0.2	49.5 ±0.6	47.0 ±0.2	48.3 ±0.2	64.0 ±0.7	**68.2 ±1.3**	67.7 ±1.0	42.2 ±0.3	66.6 ±0.8
ha-mr	44.8 ±0.6	44.7 ±0.7	45.0 ±0.3	44.5 ±0.2	51.2 ±0.3	60.7 ±1.0	**62.1 ±1.1**	38.9 ±0.5	55.9 ±1.5
ha-me	90.8 ±6.0	93.6 ±0.4	95.6 ±0.4	94.7 ±0.5	103.8 ±3.0	99.0 ±9.3	**104.8 ±0.6**	91.6 ±1.0	102.8 ±3.4
ho-m	60.4 ±3.5	74.5 ±9.1	59.1 ±3.8	67.5 ±3.8	102.3 ±0.2	40.8 ±9.9	101.7 ±0.3	65.1 ±1.6	**104.6 ±0.9**
ho-mr	64.4 ±21.5	96.4 ±5.3	95.1 ±5.3	97.4 ±6.4	95.0 ±6.5	100.3 ±0.8	99.7 ±0.8	81.8 ±6.9	**103.2 ±1.1**
ho-me	101.2 ±9.1	52.7 ±37.5	99.3 ±10.9	107.4 ±7.8	109.5 ±2.3	101.3 ±11.6	105.2 ±10.1	110.4 ±0.3	**113.3 ±1.4**
wa-m	82.7 ±4.8	66.5 ±26.0	80.8 ±3.3	80.9 ±3.2	85.8 ±0.8	87.5 ±0.7	93.4 ±1.4	67.6 ±2.5	**101.0 ±1.6**
wa-mr	85.6 ±4.0	82.2 ±1.1	73.1 ±13.2	82.2 ±3.0	84.2 ±2.3	79.0 ±0.5	87.1 ±2.8	59.9 ±2.7	**98.3 ±2.0**
wa-me	110.0 ±0.4	49.4 ±38.2	109.6 ±0.4	111.7 ±0.9	111.9 ±0.4	114.9 ±0.4	114.8 ±0.7	107.1 ±1.0	**118.1 ±0.4**
Avg.	76.5	67.7	78.3	81.6	89.7	83.5	92.9	73.8	**96.0**

- The results indicate that integrating the FANS framework into the structurally simple Actor-Critic algorithm TD3 yields the **best average performance** (Avg.) across all evaluated tasks.

Validation of FANS in Mitigating Overestimation

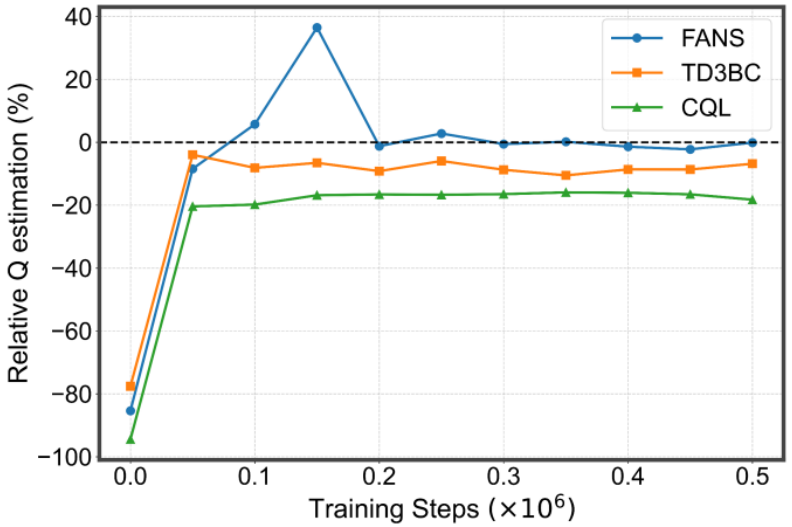
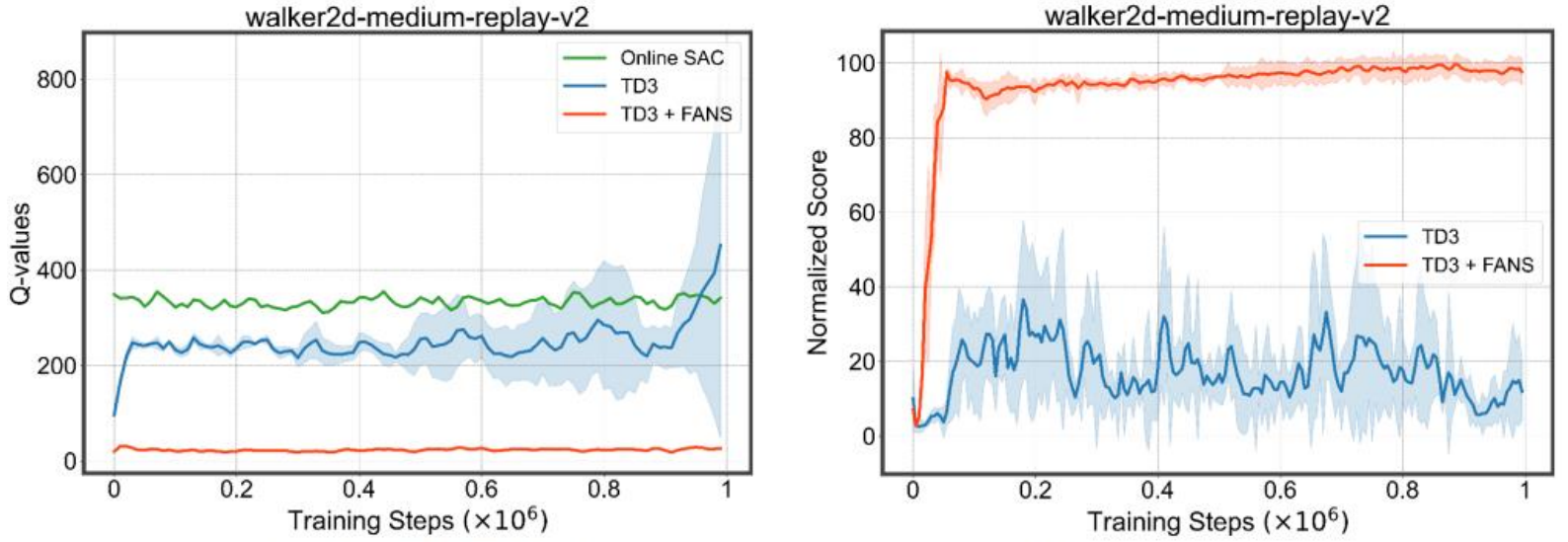


Figure 2: Relative Q estimation.



(a) Q-values for OOD data

(b) Training process

Figure 3: Evaluation of overestimation mitigation: Q-value estimates on OOD data and corresponding performance of TD3 and TD3 + FANS.

- FANS consistently achieves the **most accurate Q-value estimation.**

- FANS maintains consistently **lower Q-value estimates on OOD data**, effectively mitigating overestimation issues.

Validation of FANS in Generalization Control

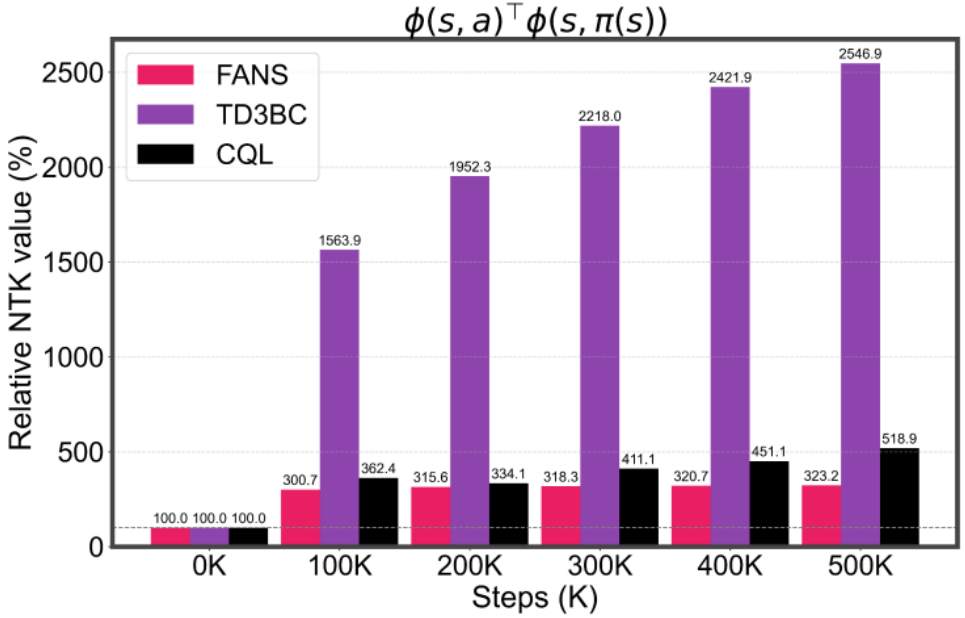


Figure 4: The NTK value of $(s, \pi(s)), s \in D$.

v.s. SOTA generalization-centric methods

Table 3: Performance comparison on D4RL locomotion tasks over the final ten evaluations and five seeds (normalized scores). We **bold** the highest mean.

Tasks	DOGE	TSRL	SPOT	POR	PRDC	STR	DIFFUSION -QL	CQL +ADS	**TD3 +FANS**
ha-m	45.3	48.2	58.4	48.8	63.5	51.8	51.1	**73.9**	66.6
ho-m	98.6	86.7	86.0	78.6	100.3	101.3	90.5	101.0	**104.6**
wa-m	86.8	77.5	86.4	81.1	85.2	85.9	87.0	91.3	**101.0**
ha-mr	42.8	42.2	52.2	43.5	55.0	47.5	47.8	49.6	**55.9**
ho-mr	76.2	78.7	100.2	98.9	100.1	100.0	101.3	102.4	**103.2**
wa-mr	87.3	66.1	91.6	76.6	92.0	85.7	95.5	93.7	**98.3**
ha-me	78.7	92.0	86.9	94.7	94.5	94.9	96.8	93.5	**102.8**
ho-me	102.7	95.9	111.4	99.3	109.2	111.9	111.1	113.3	113.3
wa-me	110.4	109.8	112.0	109.1	111.2	110.2	110.1	112.1	**118.1**
Avg.	81.0	77.5	85.9	80.1	90.1	87.7	87.9	92.3	**96.0**

- The significantly lower NTK values in FANS demonstrate a fundamental **suppression of pathological generalization** patterns.
- Among the many advanced methods associated with improving generalization, FANS remains **highly competitive**, achieving the best average performance.

Ablation Study

Table 4: Ablation study of FANS. M is the number of ensemble in FANS.

Tasks	TD3+FANS w/o Residual	TD3+FANS w/o Gaussian	TD3+FANS w/o LayNorm	TD3+FANS w/o Ensemble	**TD3 + FANS**	
					Score	M
ha-m	66.4 ± 1.3	66.2 ± 0.8	64.3 ± 3.5	66.6 ± 0.8	$\mathbf{66.6 \pm 0.8}$	2
ho-m	21.0 ± 19.3	62.9 ± 30.5	99.6 ± 6.9	79.6 ± 26.6	$\mathbf{104.6 \pm 0.9}$	5
wa-m	6.5 ± 4.3	7.7 ± 7.8	91.9 ± 2.3	6.7 ± 0.4	$\mathbf{101.0 \pm 1.6}$	5
ha-mr	55.1 ± 0.5	54.4 ± 1.2	55.6 ± 1.3	55.1 ± 1.3	$\mathbf{55.9 \pm 1.5}$	3
ho-mr	41.7 ± 7.2	95.1 ± 4.5	99.2 ± 5.6	44.6 ± 6.3	$\mathbf{103.2 \pm 1.1}$	3
wa-mr	35.6 ± 11.3	80.2 ± 22.1	95.1 ± 1.7	75.4 ± 23.5	$\mathbf{98.3 \pm 2.0}$	3
ha-me	64.5 ± 21.1	102.6 ± 3.1	51.1 ± 4.7	28.4 ± 1.4	$\mathbf{102.8 \pm 3.4}$	5
ho-me	1.5 ± 0.3	30.3 ± 8.2	41.3 ± 13.8	1.6 ± 0.8	$\mathbf{113.3 \pm 1.4}$	5
wa-me	-0.2 ± 0.1	89.1 ± 21.0	115.2 ± 0.4	16.5 ± 10.7	$\mathbf{118.1 \pm 0.4}$	5
Avg.	32.3	65.4	79.3	41.6	**96.0**	-

- **Every module contributes to** the overall performance, though to varying degrees.

Table 5: Ablation study on residual placement: performance drops when residual blocks are added to the actor network.

Structure	**Control Setup**	**TD3 + FANS**
Actor + Residual	✓	×
Critic + Residual	✓	✓
ha-m	63.8 ± 0.1	$\mathbf{66.6 \pm 0.8}$
ho-m	21.8 ± 19.8	$\mathbf{104.6 \pm 0.9}$
wa-m	50.3 ± 35.2	$\mathbf{101.0 \pm 1.6}$

- Applying residual connections exclusively in the **critic network** facilitating stable value estimation without constraining the action space, thereby **achieving better overall performance**.

Conclusion

- **A novel network architecture framework designed to tackle the unique generalization challenges of offline RL**

  By integrating residual blocks, Gaussian activation functions, layer normalization, and model ensembling, FANS systematically steers optimization toward flatter minima, thereby enhancing stability and reducing overfitting.

- **Comprehensive analyses**

  Our comprehensive analyses reveal the individual and combined effects of these components in promoting smoother optimization landscapes and lowering variance.

- **Contributions to the community**

  Our work highlight the critical role of architectural design as a complementary and effective approach to advancing offline RL performance, opening promising avenues for future research.


SHANXI UNIVERSITY
1902
NORTH UNIVERSITY OF CHINA
1941
TIANJIN UNIVERSITY (PEIYANG UNIVERSITY)
1895

Thanks