Neural Processes with Stochastic Attention

: Paying more attention to the context dataset

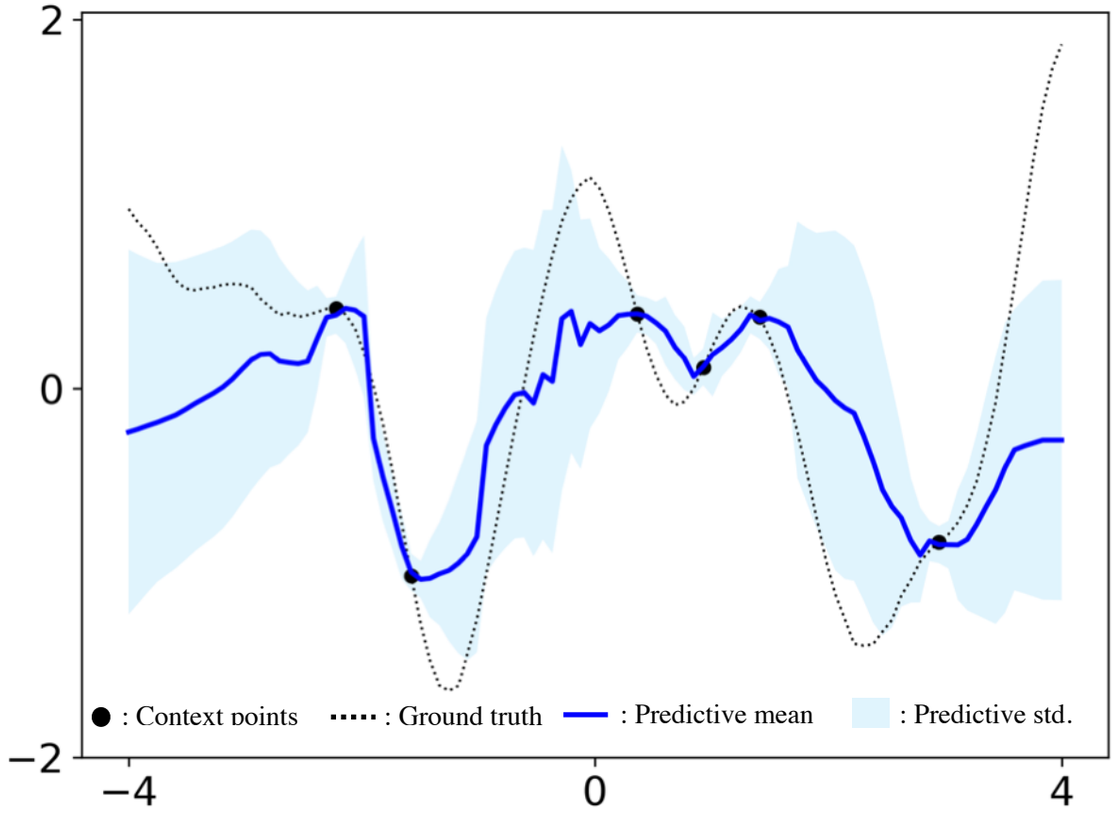
Mingyu Kim, Kyeongryeol Go and Se-young Yun

Optimization and Statistical Inference LAB

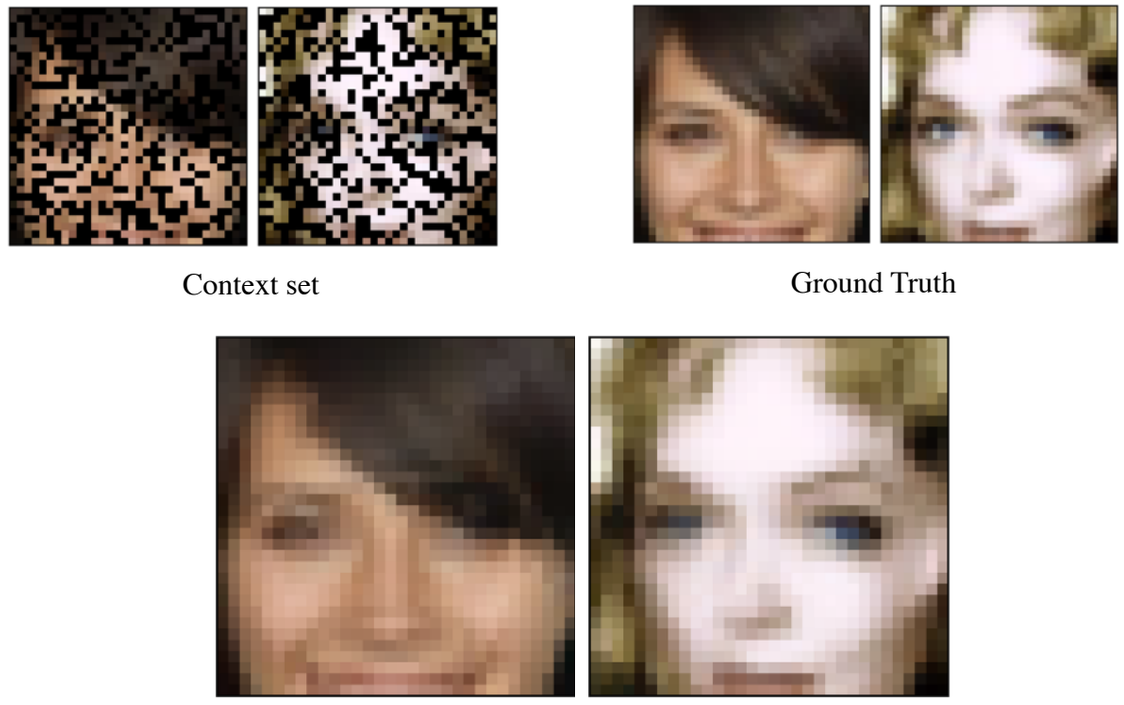
KAIST AI
Kim Jaechul Graduate School

Neural Processes (NPs)

Optimization and Statistical Inference LAB



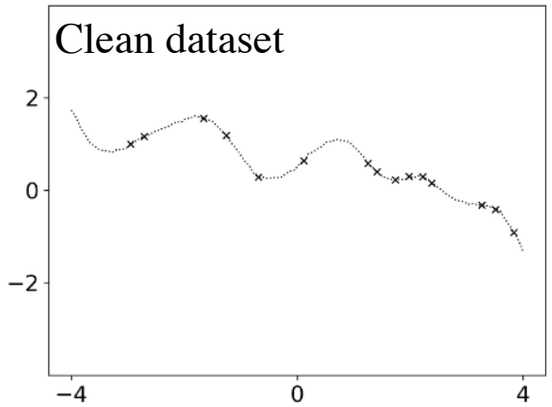
(a) 1D regression regression

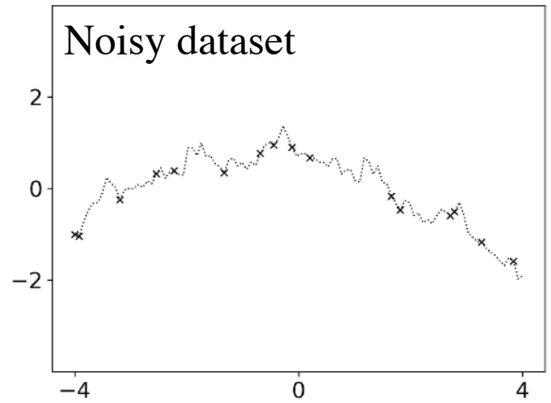


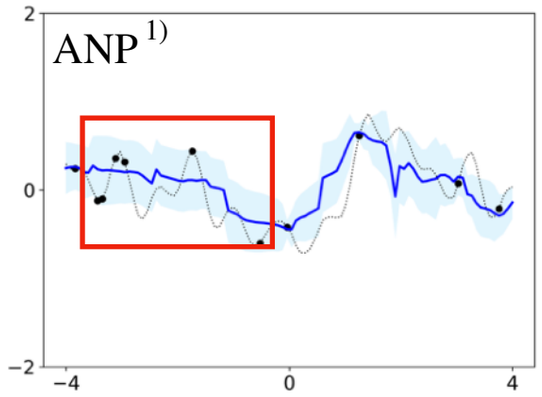
(b) Image completion

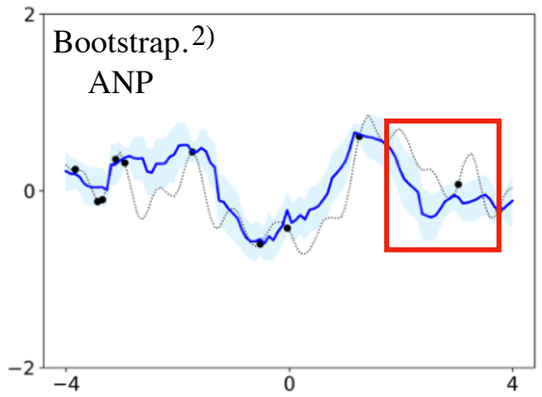
- Neural Processes implicitly describe a large class of stochastic processes with neural networks.
- It complete unseen target points considering a given context dataset without huge computation.
- Attention mechanism is mostly used for context set encoding in terms of performance.

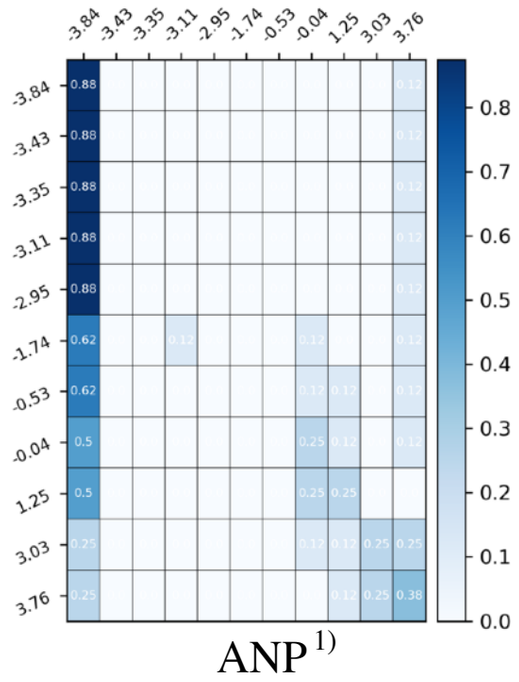
Noisy situation

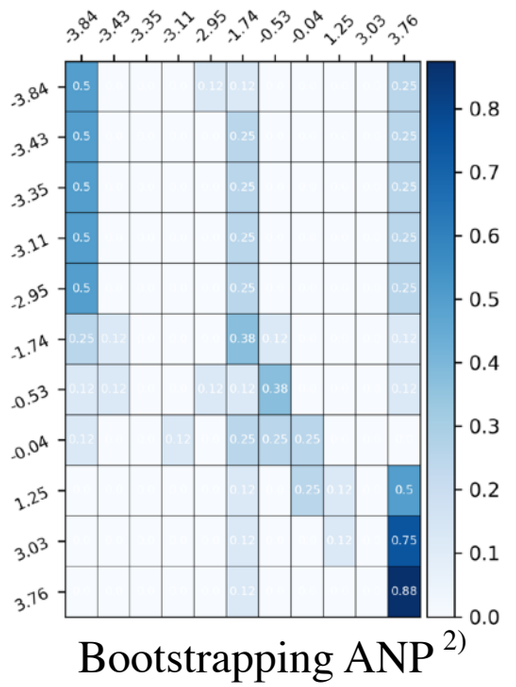












(a) Datasets (b) Prediction results (c) Attention heat-map results

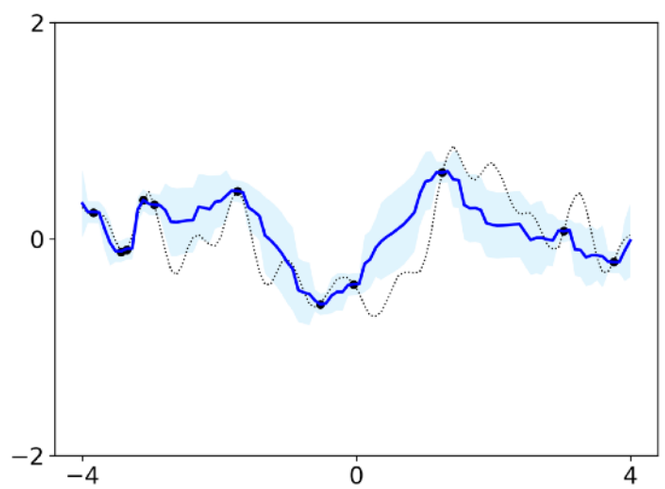
- While existing NPs do well in terms of prediction, they do not properly capture context data points.
- When we analyze attention heat-maps, we identify that the baselines are far from the ideal case. The best pattern is diagonal because all feature values are arranged in ascending order.

1) Kim, Hyunjik, et al. "Attentive Neural Processes." *International Conference on Learning Representations*. 2019.
2) Lee, Juho, et al. "Bootstrapping neural processes." *Advances in neural information processing systems*. 2020.

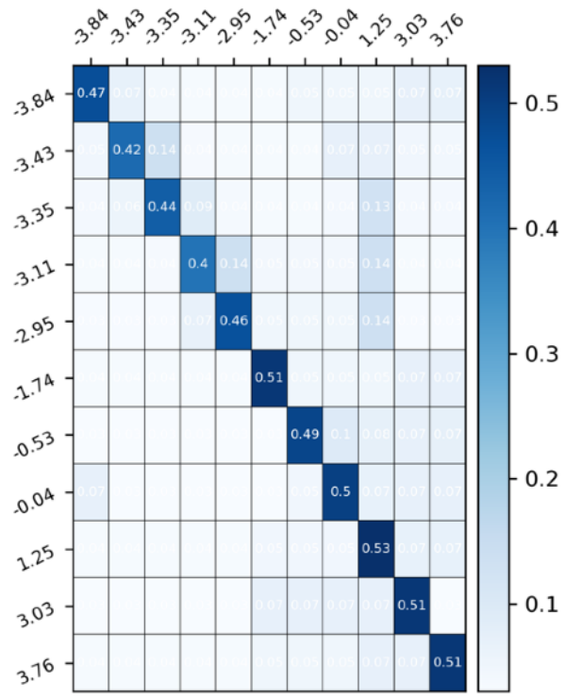
Neural Processes with Stochastic Attention

Optimization and Statistical Inference LAB

- For stochastic attention, we employ Bayesian Attention Modules, which enable completely amortized variational inference.
- We claims for the first time, using the information theory framework, that critical conditions for contextual embeddings in NPs are independent of target features and close to contextual datasets.
- We show that proposed model substantially outperforms conventional NPs in typical meta-regression problems.



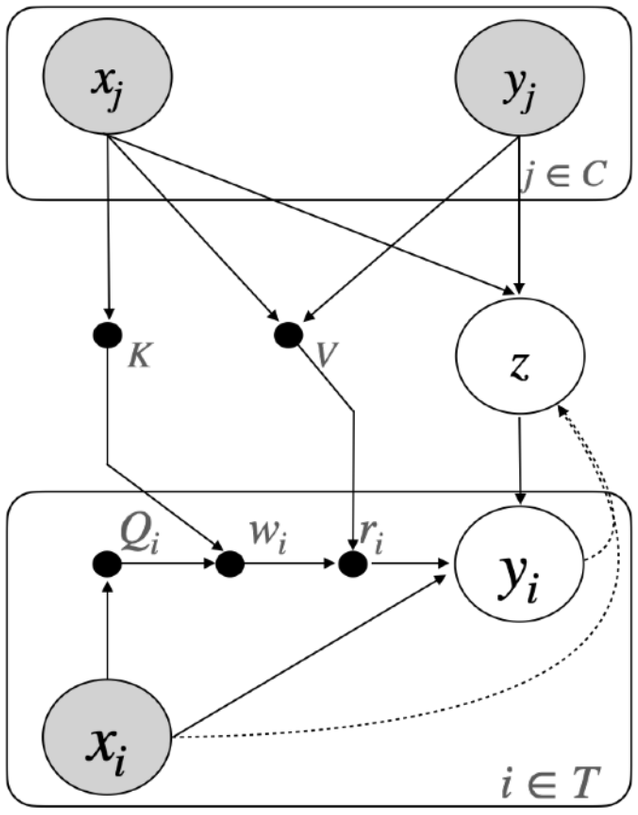
(a) 1D regression by ours



(b) Heat-map by ours

Neural Processes with Stochastic Attention (1)

Optimization and Statistical Inference LAB



(a) ANP

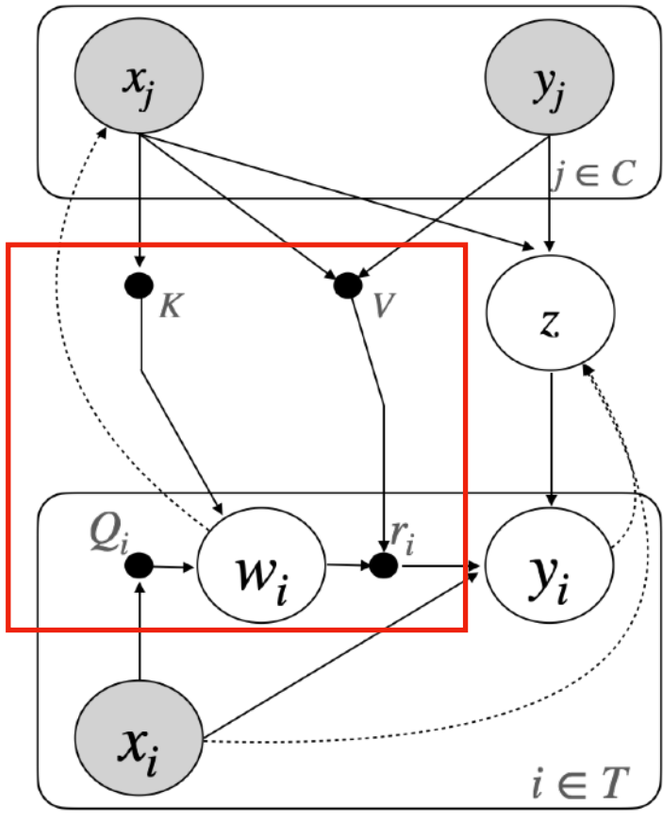
$$w_{i,j} = \text{softmax}\left(\frac{q_i^T k_j}{\sqrt{d_k}}\right)$$

● : Deterministic variables

◯ : Stochastic variables

——— : Generation

- - - - - : Inference



(b) Ours

- Attention weights sampled from the Weibull distribution, satisfying non-negative values, $q(w_i \mid x_i, X_c)$.
- The prior distribution is the Gamma distribution using only the context dataset, $q(w_i \mid X_c)$

✻ key contextual prior

- The attention weights strongly preserve significant amounts of context information by decreasing $KL[q(w_i \mid x_i, X_c) \mid q(w_i \mid X_c)]$

3) Fan, Xinjie, et al. "Bayesian attention modules." *Advances in Neural Information Processing Systems*. 2020

Neural Processes with Stochastic Attention (1)

Optimization and Statistical Inference LAB

- As we reveal that (1) is bounded, we achieve that NPs learn contextual embeddings that are independent of target features (noisy situation) and close to ideal case.

Theorem 1.

(1)
$$\mathscr{L}_{\mathscr{T}_k}(\phi,\theta)=\sum_{i=1}^{N}\left[\log p_\theta(y_i\,|\,x_i,z,r_i)-\mathsf{KL}\left(q_\phi(w_i\,|\,x_i,X_c)\,|\,q_\phi(w_i\,|\,X_c)\right)\right]-\mathsf{KL}\left(q_\phi(z\,|\,X,Y)\,|\,q_\phi(z\,|\,X_c,Y_c)\right)$$

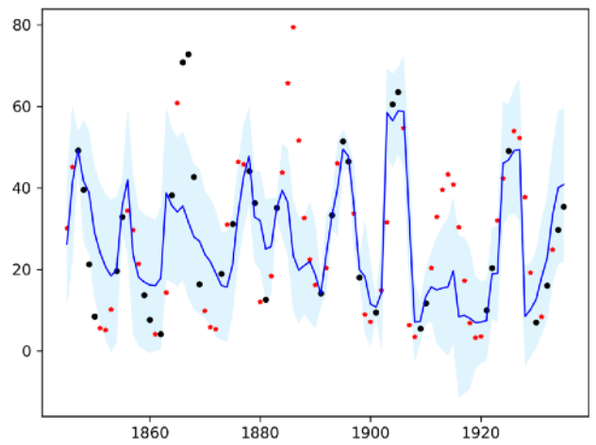
(2)
$$\mathscr{L}_{\mathscr{T}_k}(\phi,\theta)\le I(y_i;\mathscr{D}\,|\,x_i)-I(Z,x_i\,|\,\mathscr{D})$$

where, Z means representation of context dataset and $\mathscr{D}$ is a context dataset

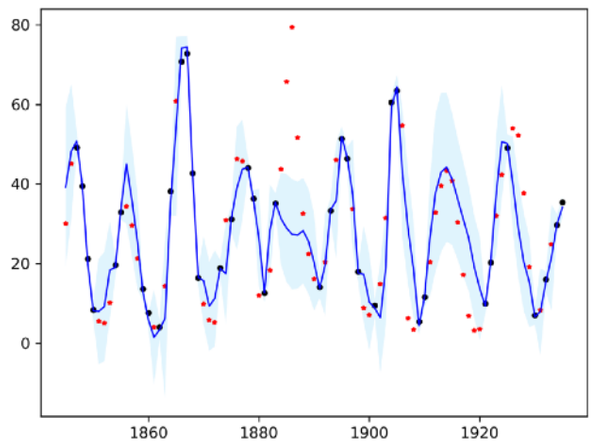
- We maximize (1) that the typical NPs objective function and KLD of Stochastic attentions.
  - $I(y_i;\mathscr{D}\,|\,x_i)$: Measure to identify that NPs is adapted to a novel task.
  - $I(Z,x_i\,|\,\mathscr{D})$: If context representation Z have totally different information against the target feature x_i given the context dataset $\mathscr{D}$, It can be 0.

Experiments : Sim2Real

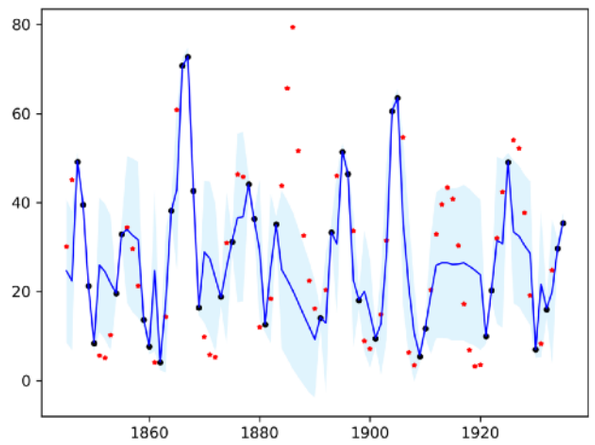
Optimization and Statistical Inference LAB



(a) Bootstrapping ANP



(b) ConvCNP



(c) Ours

* These figures show prediction results when all models trained on noisy situation

	Clean situation		**Noisy situation**	
	context	target	context	Target
ANP	-1.756	-3.742	-0.634	-1.962
Bootstrapping ANP	**2.451**	-3.382	-1.183	-2.008
ConvCNP [4)]	1.758	**-0.431**	1.879	-0.205
Ours	**2.429**	-1.766	**2.699**	**-0.076**

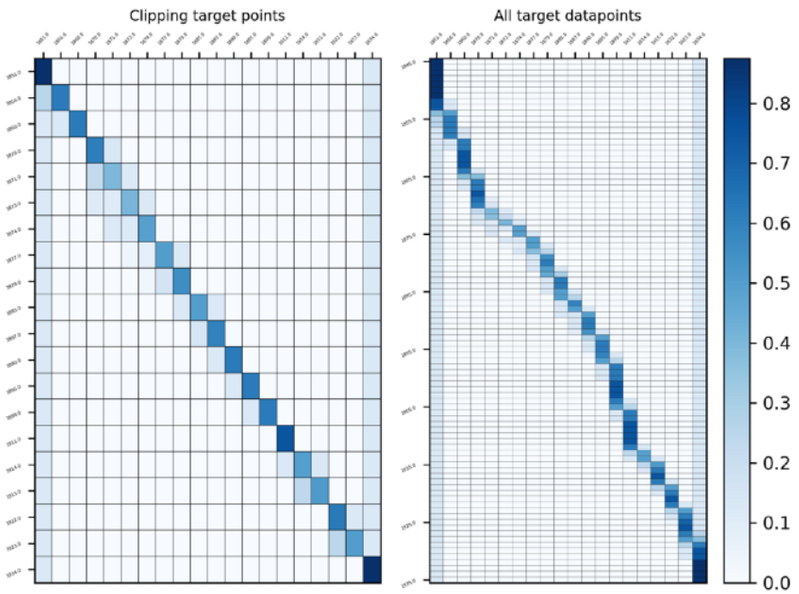
Metric : Loglikelihood

4) Gordon, Jonathan, et al. "Convolutional Conditional Neural Processes." *International Conference on Learning Representations*. 2020.

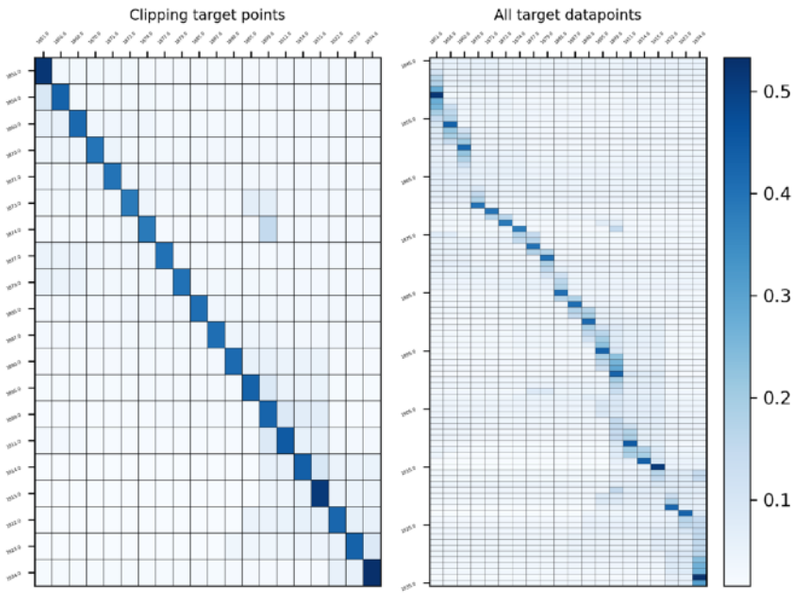
Experiments : Sim2Real

Optimization and Statistical Inference LAB

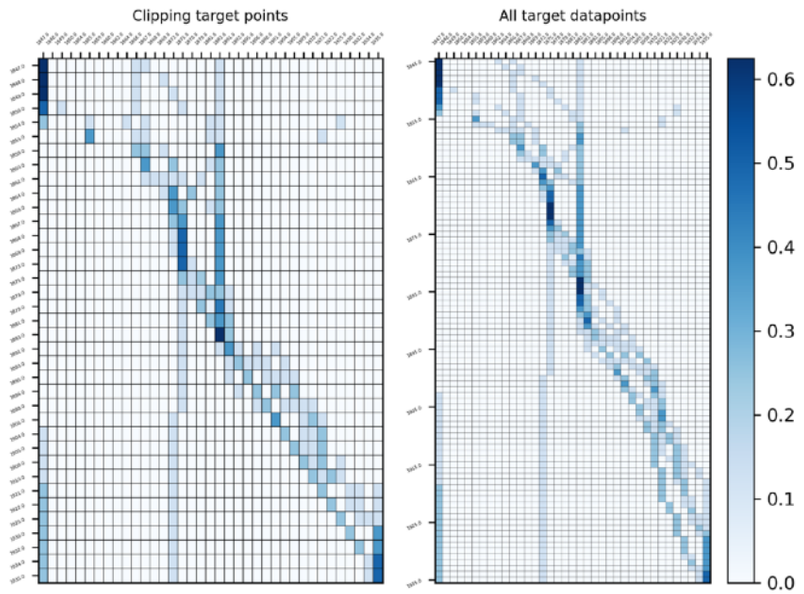
Trained on a clean dataset



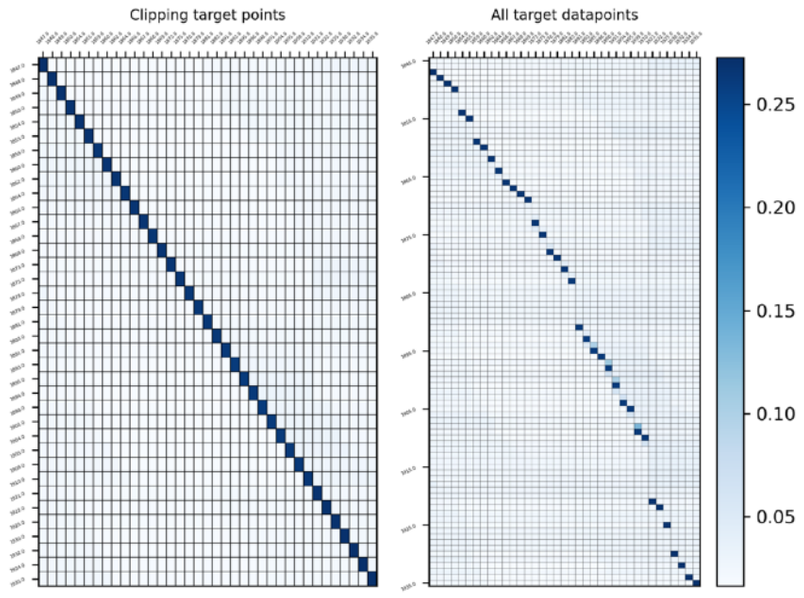
(a) Bootstrapping ANP



Trained on a noisy dataset



(b) Ours



Experiments : Image Completion

Optimization and Statistical Inference LAB

Ground Truth

Contexts 300

ANP

Bootstrapping ANP

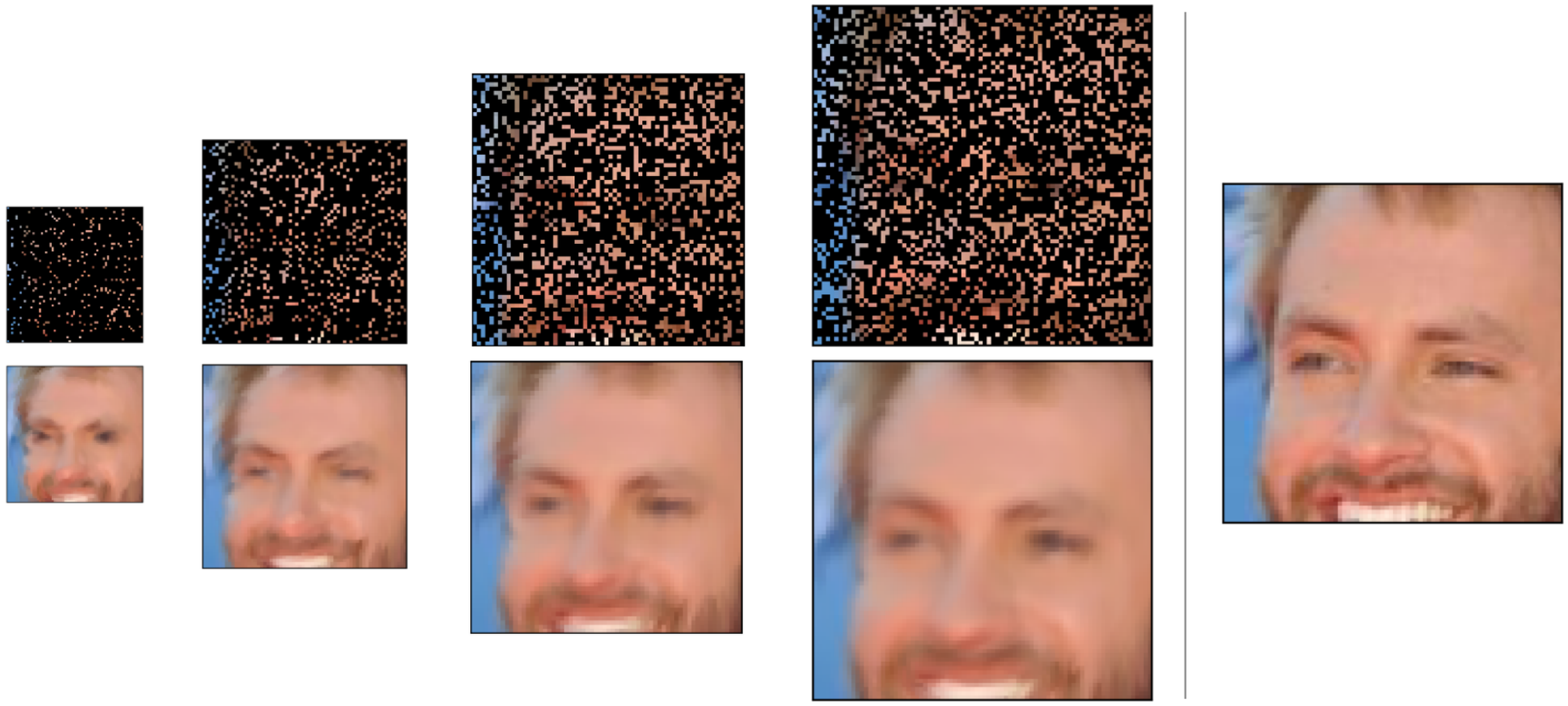
Ours

- Our model accurate capture the form of colored hair, but other baselines do not.
- Our model outperforms other baselines in terms of quantitative analysis.

	Context	**Target**			
		Context : 50	Context : 100	Context : 300	Context : 500
ANP	3.100	2.492	2.806	3.02	3.06
Bootstrapping ANP	**3.172**	2.453	2.837	3.095	3.145
Ours	**4.119**	2.653	**3.21**	**3.787**	**3.948**

Experiments : Image Completion

Optimization and Statistical Inference LAB



32 × 32 48 × 48 64 × 64 80 × 80 Ground Truth

- The training dataset consists of the CelebA images with a size of 32 × 32. The completed images range in size from 32 × 32 to 80 × 80. Each image is completed with 0.3 of the overall pixels.
- We identify that ours completes images that are higher resolution than the training images.

Conclusion

- We claims the conditions for contextual embeddings in NPs are independent of target features and close to contextual datasets via information theory.

- The Neural processes with stochastic attention outperform previous baselines

- We reveal that our attention weights of our model are more stable than previous model in noisy situations.