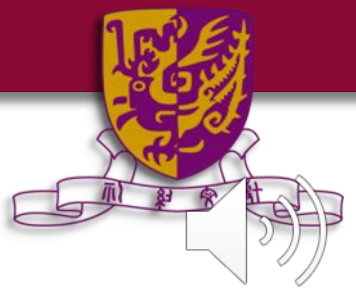
Attentive Normalization for Conditional Image generation

Yi Wang[1], Ying-Cong Chen[1], Xiangyu Zhang[2], Jian Sun[2], Jiaya Jia[1]

[1] The Chinese University of Hong Kong

[2] MEGVII Technology



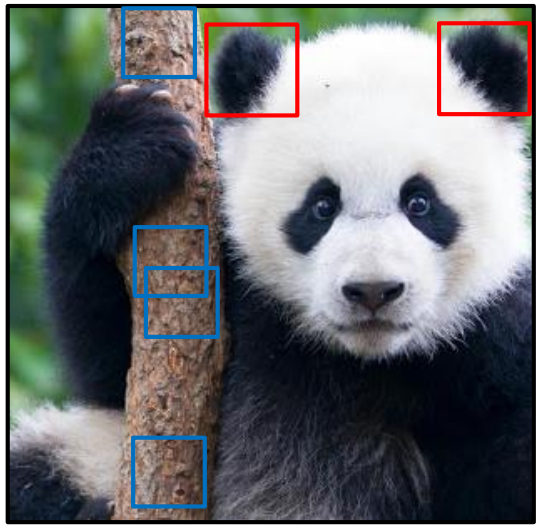
What is Conditional Image generation?

panda
Input Class Label

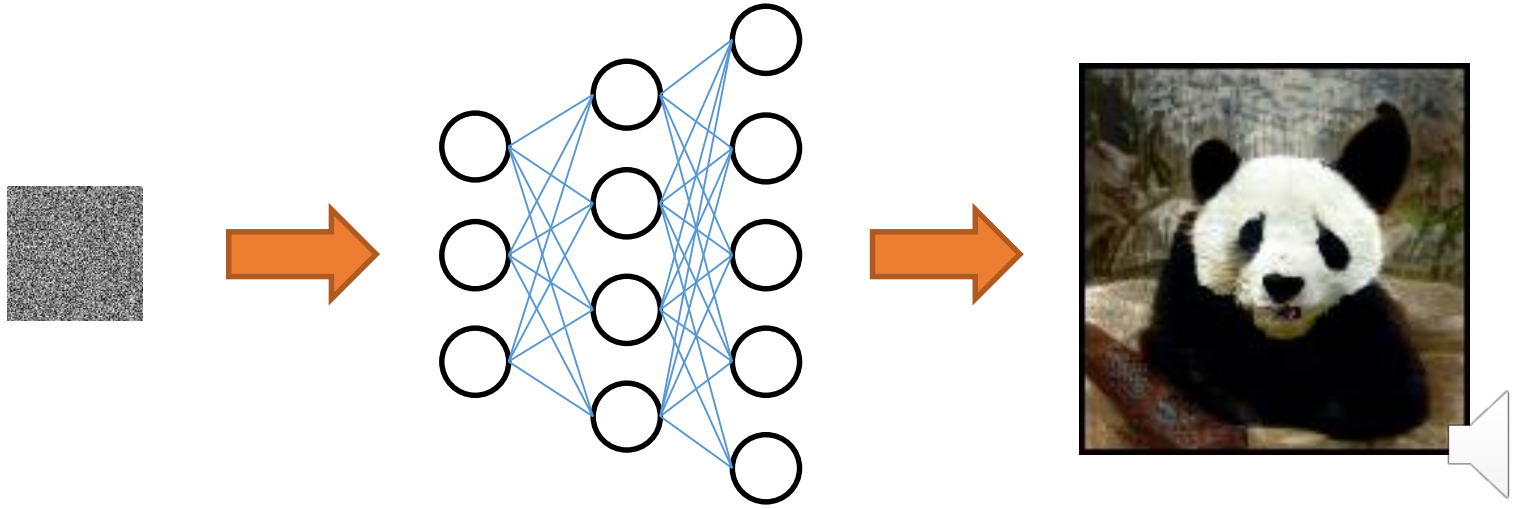
Generation

Generative model

Long-range dependency in image generation

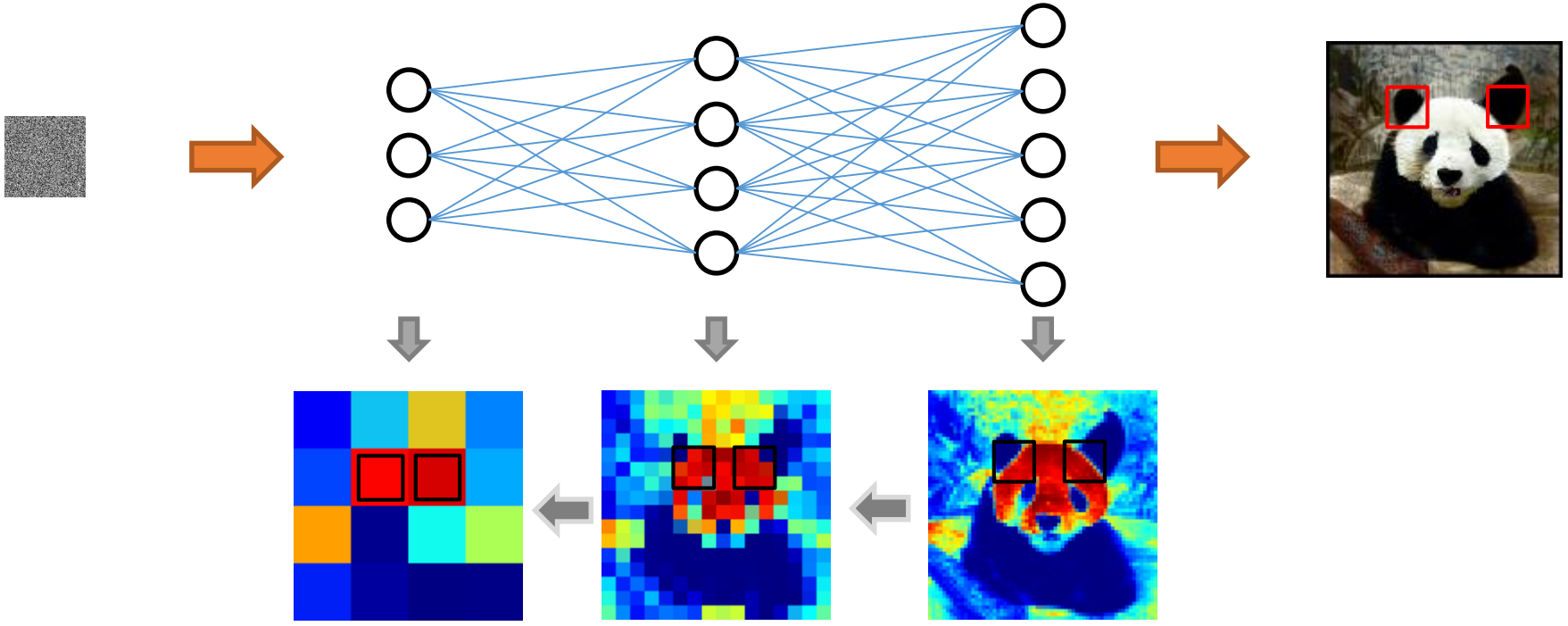


- **Standard convolutional neural network:**
  - **Modelling image contents in a hierarchical manner.**

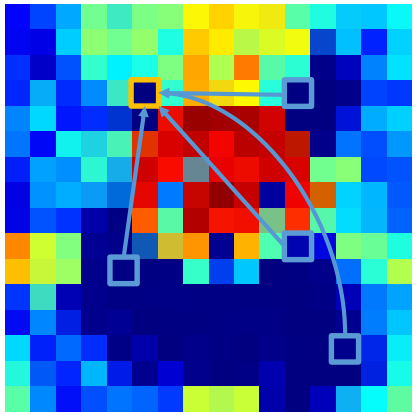


Long-range dependency in image generation

- **Standard convolutional neural network:**
  - **Long-range dependency is conduct in a Markov chain.**



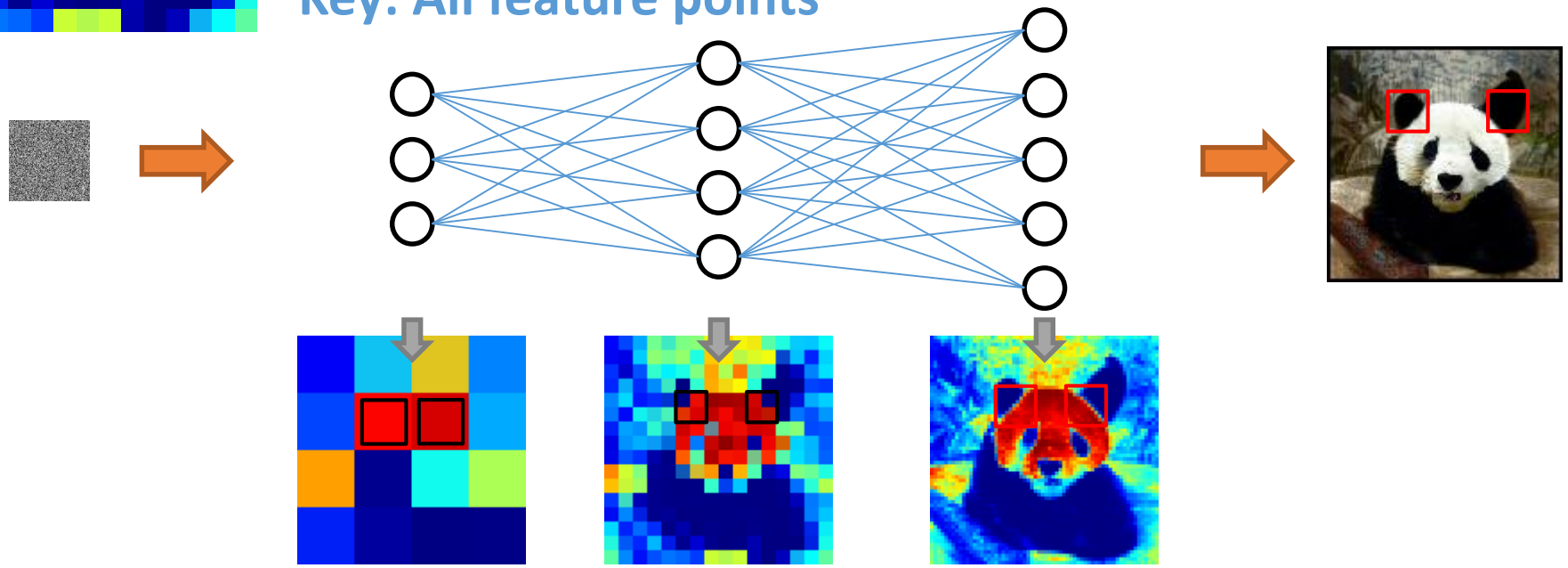
Prior work: self-attention



Self attention: reconstructing each feature point using the weighted sum of all feature points[1].

Query: every feature point

Key: All feature points

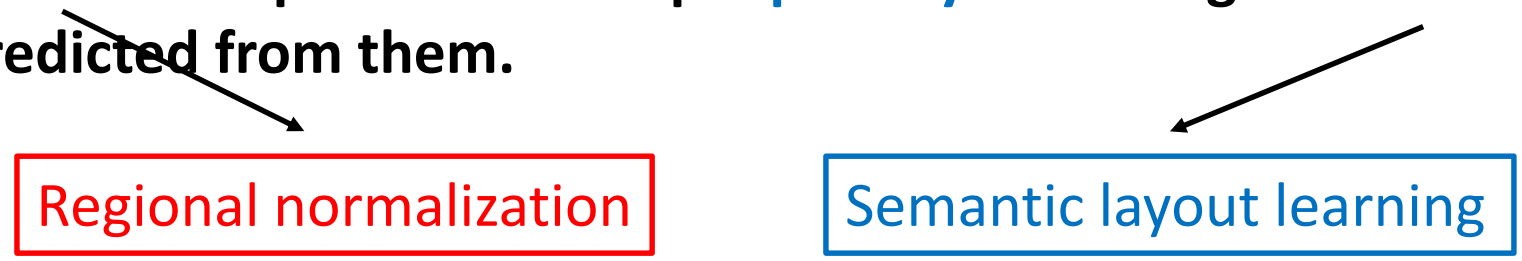


[1] Zhang, Han, et al. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

Our method: Attentive Normalization

Core idea:

We normalize the input feature maps spatially according to the semantic layouts predicted from them.



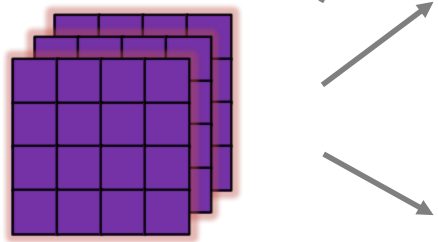
Empirical observations to backup our method:

- **A feature map can be viewed as a composition of multiple semantic entities[3,4].**
- **The deep layers in a neural network capture high-level semantics of the input images[5].**

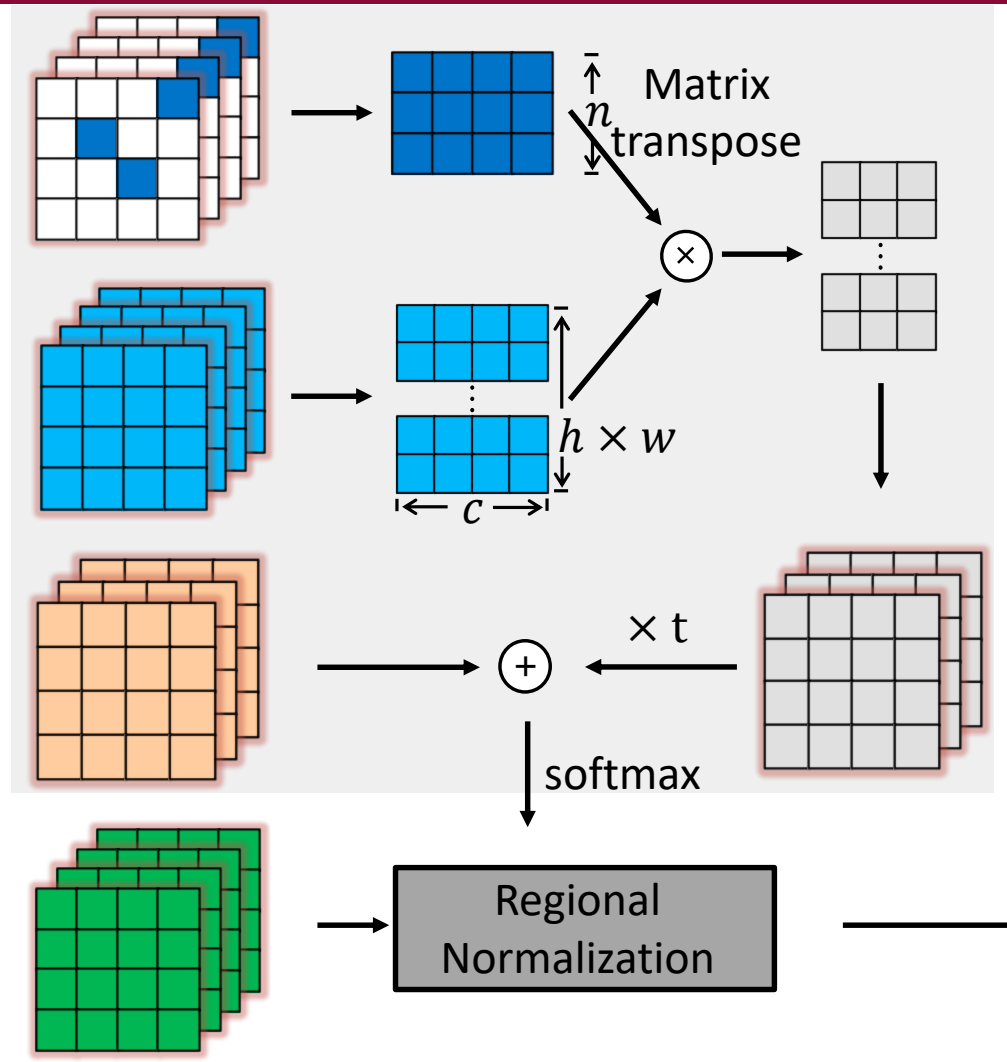
[3] Greff, Klaus, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. In *NeurIPS*, 2017.
[4] Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *NeurIPS*, 2017.
[5] Le, Quoc V. Building high-level features using large scale unsupervised learning. In *ICASSP*, 2013.

Our method: Attentive Normalization

Semantic layout learning
+
Regional Normalization



Input feature maps X



Forward flow

Convolution

Matrix multiplication

Semantic layout learning

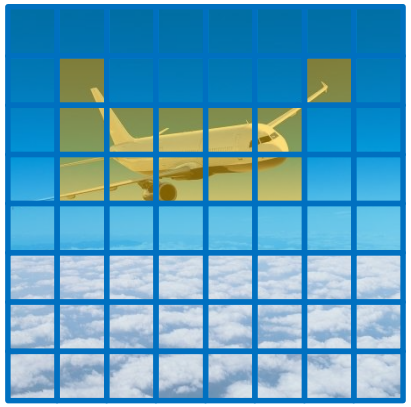
Semantic layout learning

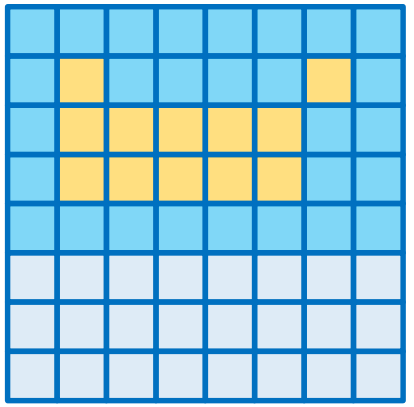
An image is composed of n semantic entities.

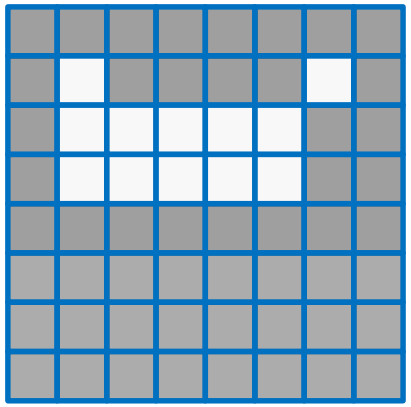
Each feature point of the image, it is determined by at least one entity.

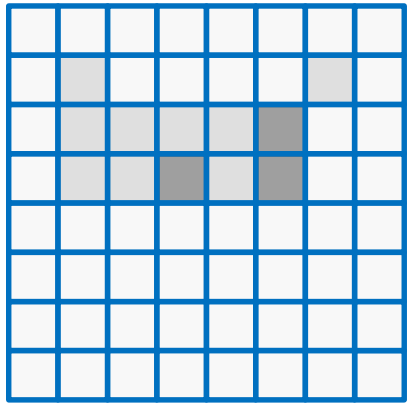
Semantic layout learning

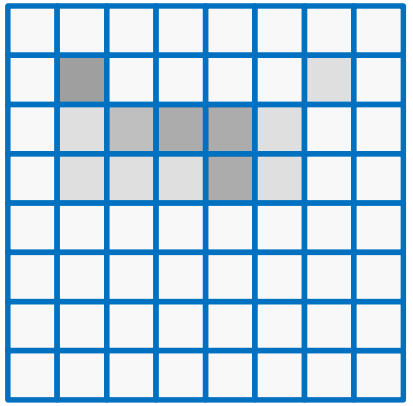
How to group feature points of an image according to their correlation to the semantic entities.

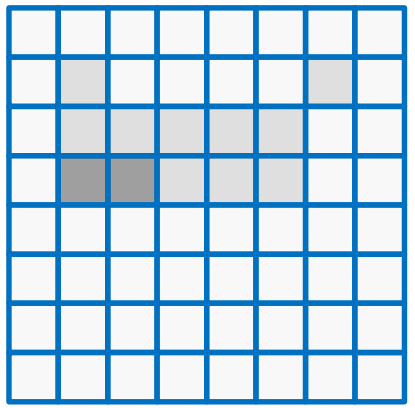












Sky
Plane
Cloud

Learned entities
Background
Round thing
Square thing
Flat thing

Semantic layout learning

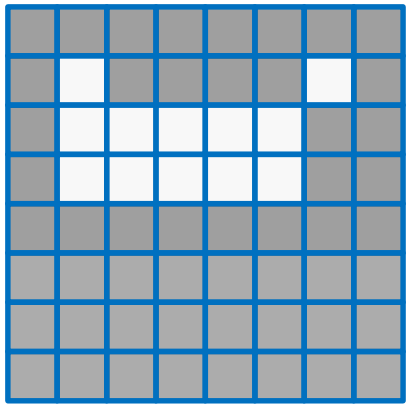
How to get these semantic entities?

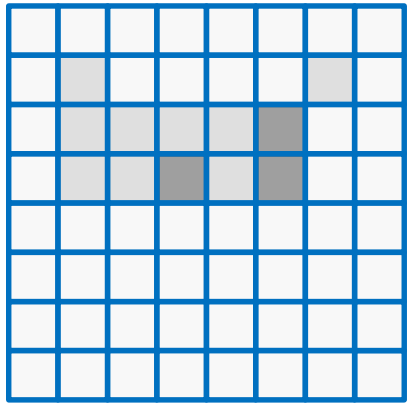
Learned entities

- **Background**
- **Round thing**
- **• • •**
- **Square thing**
- **Flat thing**

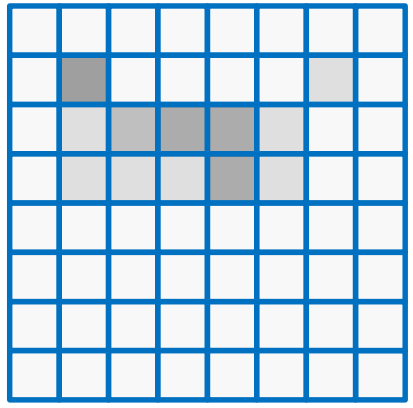
Implementations

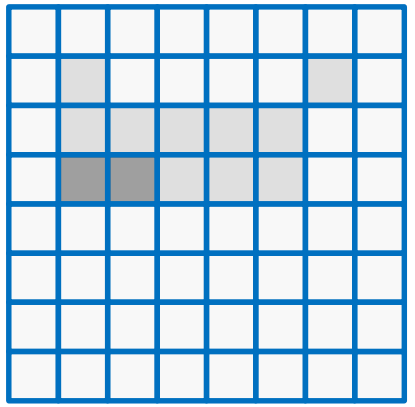
- **A convolutional layer with n filters (**n denotes a predefined class number**)**





• • •

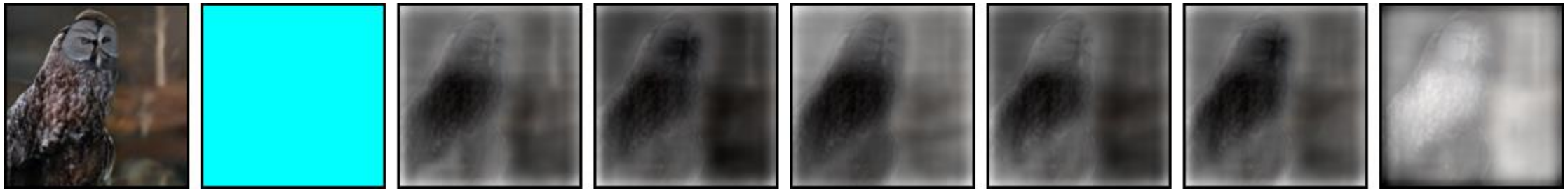




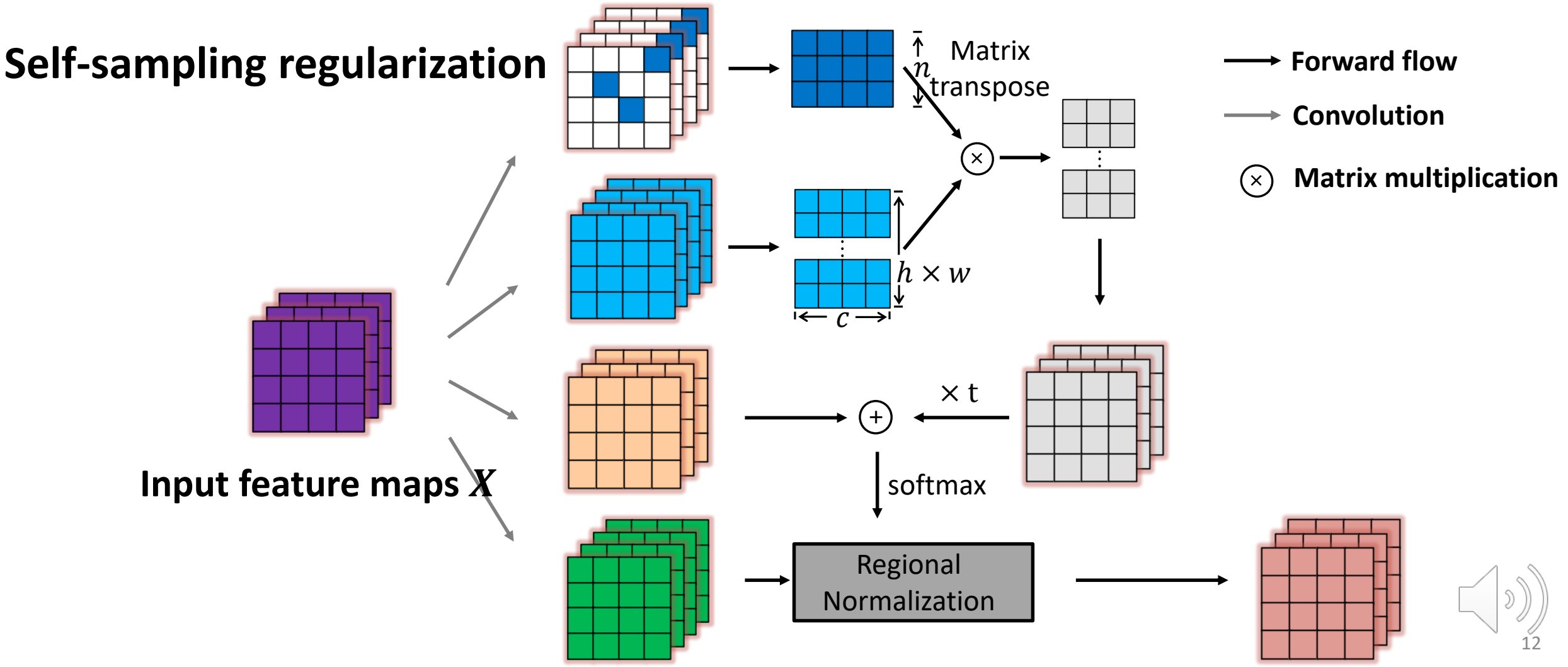
Challenges in optimizing SSL

Trivial solutions for learning entities directly

- It tends to group all feature points with a single semantic entities.
- No protocols are set to ban useless semantic entities.

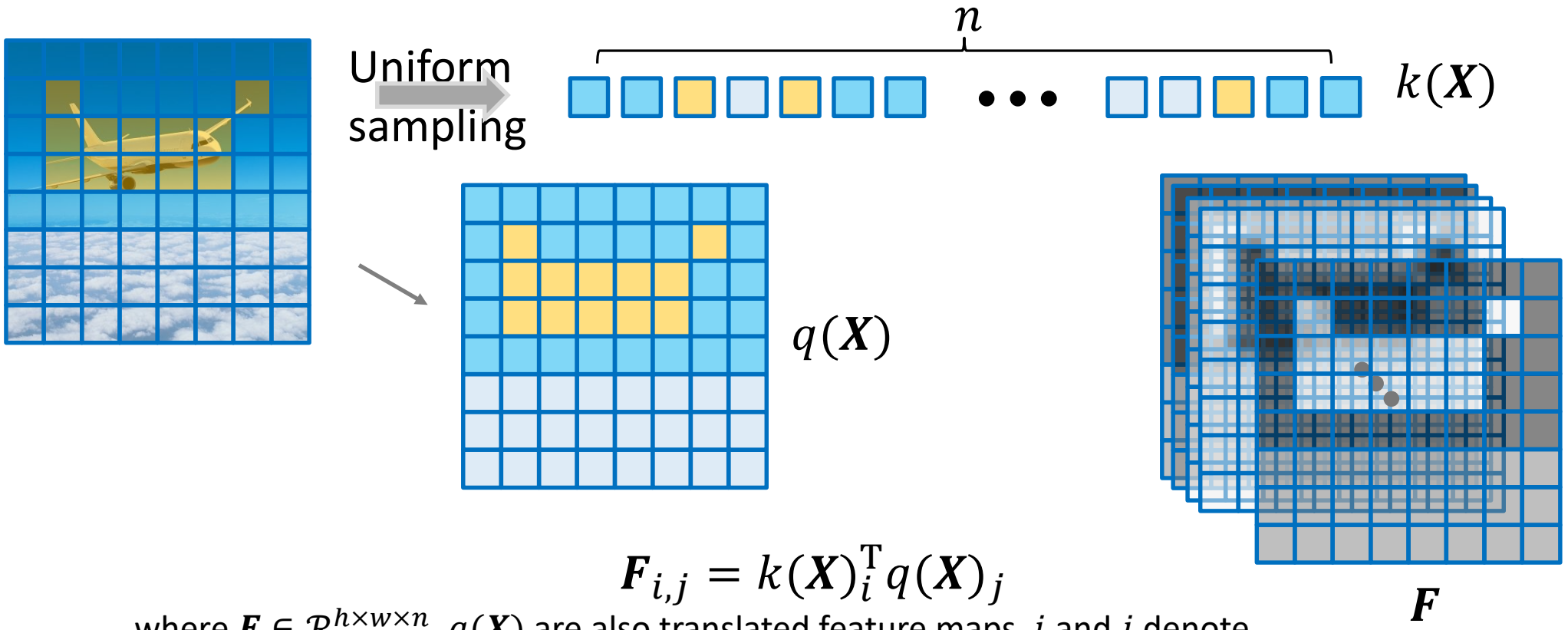


Our method: Attentive Normalization



Self-Sampling Regularization (SSR)

Regularizing semantics learning with a self-sampling branch



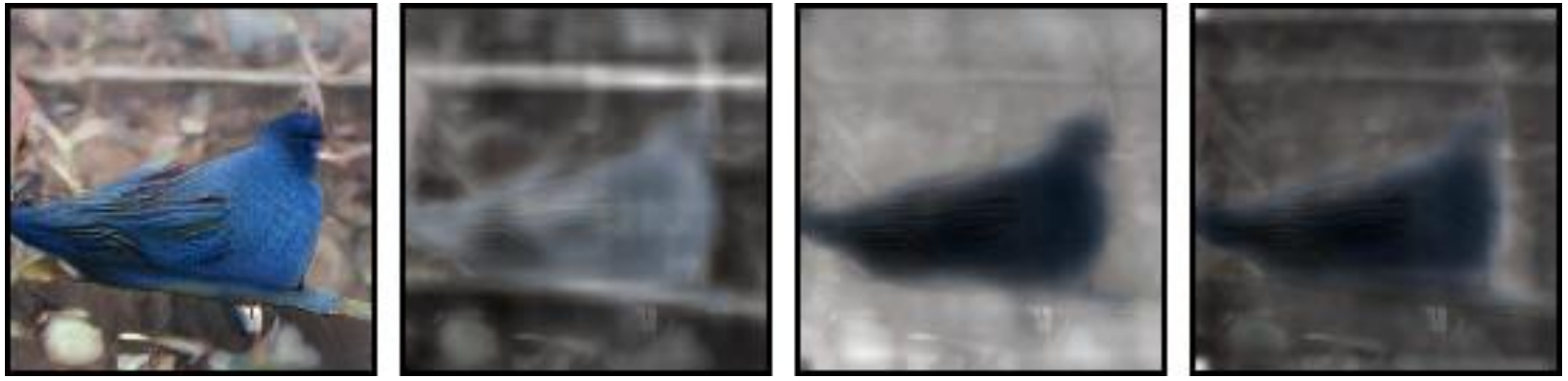
$$\boldsymbol{F}_{i,j} = k(\boldsymbol{X})_i^{\mathrm{T}} q(\boldsymbol{X})_j$$

where $\boldsymbol{F} \in \mathcal{R}^{h \times w \times n}$, $q(\boldsymbol{X})$ are also translated feature maps. i and j denote pixel location. We set $\#\{i\} = n$ and $\#\{j\} = h \times w$.

Analysis: learned semantic layouts

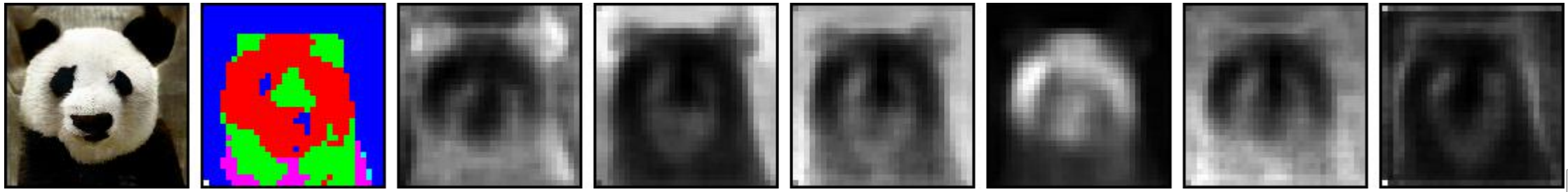
The predicted semantic layout indicates regions with high inner coherence in semantics.

(-0.01, 4.55) (0.03, 0.68) (0.03, 0.80) (-0.03, 1.07)



Mean

Standard deviation



Generated

The highlighted regions indicated by the learned semantic layouts

Analysis: complexity analysis

Complexity Analysis

The computational complexity of Attentive Normalization: $O(nNHWC)$

The computational complexity of Self-Attention: $O(N(H^2W^2C + HWC^2))$

Module (ms)	128 x 128	256 x 256	512 x 512	1024 x 1024
AN (n=16)	**0.73**	**2.24**	**9.46**	**37.68**
Self-attention[1]	5.21	79.42	-	-

All fed tensors are with the same batch size 1 and channel number 32. Resolutions are different.

'-' stands for evaluation time unmeasurable due to out-of-memory in GPU.

Running environment: Pytorch 1.1.0, 4 CPUs, 1 TiTAN 2080 GPU, 32GB Memory.

[1] Zhang, Han, et al. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

Quantitative Results

Class-conditional image generation On ImageNet (128x128):

	Itr x 1K↓	FID↓	Intra FID↓	IS↑
AC-GAN[6]	/	/	260.0	28.5
SN-GAN[7]	1000	27.62	92.4	36.80
SN-GAN*[1]	1000	22.96	/	42.87
SA-GAN[1]	1000	18.65	83.7	**52.52**
Ours	**880**	**17.84**	**83.4**	46.57

[1] Zhang, Han, et al. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.
[6] Odena, Augustus, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017.
[7] Miyato, Takeru, and Masanori Koyama. cGANs with projection discriminator. *arXiv preprint arXiv:1802.05637*. 2018.

Qualitative Results

Class-conditional image generation


Drilling platform (540)


Agaric (992)


Schooner (780)

Image inpainting

Categorical interpolation

Blenheim spaniel<->indigo hunting<->schooner


coffee<->owl


Panda<->Drilling platform

Flower<->bird

Thanks