Food Image Synthesis from Ingredients

Fangda Han, Rutgers University

2nd Workshop on AIxFood

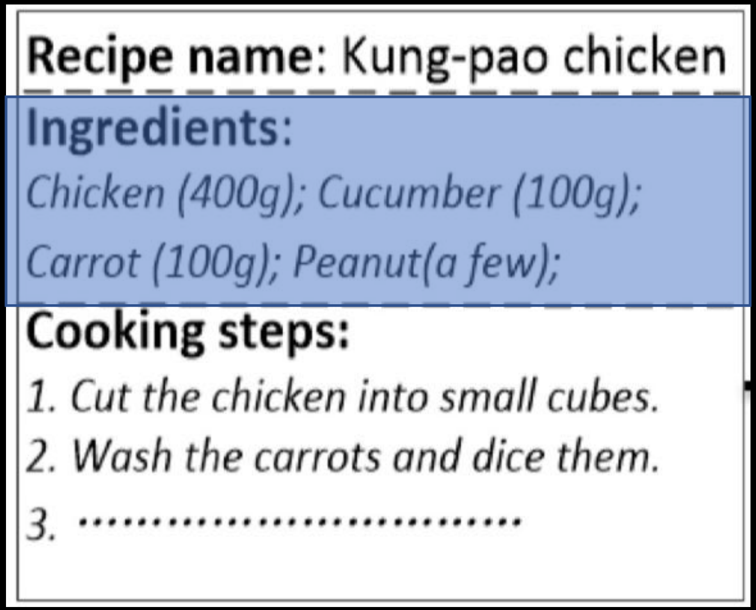
IJCAI 2020

1/7/2021

Agenda

- A Food Image Generator based on StackGAN-v2
  - Methodology
  - Results
  - Demo
- A Multi-ingredients Pizza Generator based on StyleGAN2
  - Methodology
  - Results
  - Demo

Introduction


Recipe name: Kung-pao chicken

Ingredients:

Chicken (400g); Cucumber (100g);
Carrot (100g); Peanut(a few);

Cooking steps:

1. Cut the chicken into small cubes.
2. Wash the carrots and dice them.
3. ....

Motivation

egg, salt, potato, cheese

Data Augmentation

Gamification

Food Art

Motivation

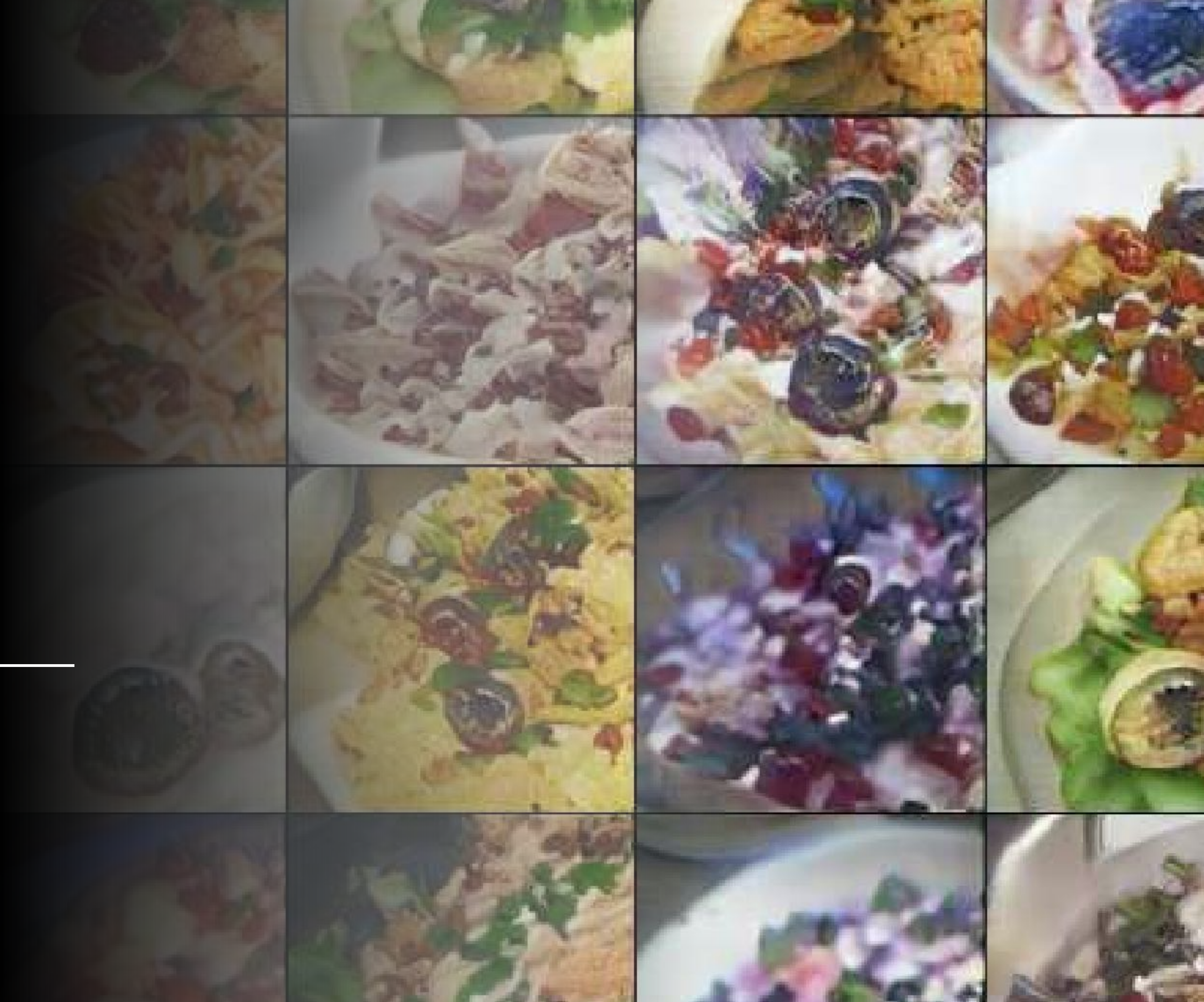
Visualize Menu

Helps young children learn to think and express ideas*

Design Cuisine

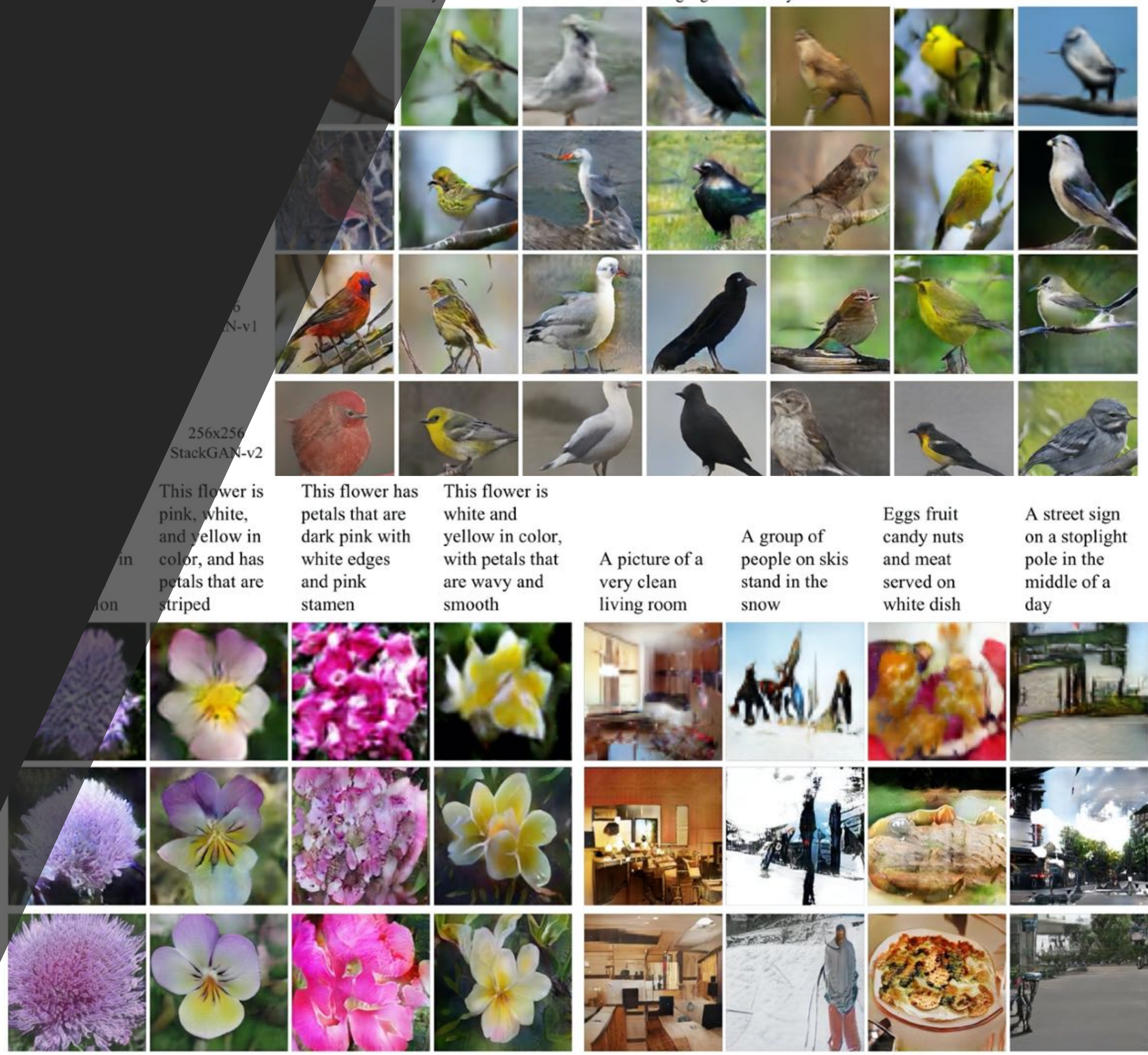
- https://edsource.org/2015/art-appreciation-helps-young-children-learn-to-think-and-express-ideas/77734
- https://www.fiverr.com/graphixerlk/create-a-professional-restaurant-menu-design
- https://www.forbes.com/sites/lisakocay/2017/08/11/food-wines-best-new-chef-2016-will-open-two-restaurants-in-miami-design-district/?sh=5670d57e3baf

A Food Image Generator based on StackGAN-v2



About StackGAN-v2

- StackGAN-v2 can generate photo-realistic images
- Works well on spatially compact objects such as birds and flowers
- Not so good on more complex scenes or objects like food
- Why?
  - More than one object
  - Color diversity
  - Interaction between objects



g. 4: Example results by our StackGANs and GAN-INT-CLS [35] conditioned on text descriptions from Oxford-102 test set (leftmost four olumns) and COCO validation set (rightmost four columns).

A Food Image Generator based on StackGAN-v2

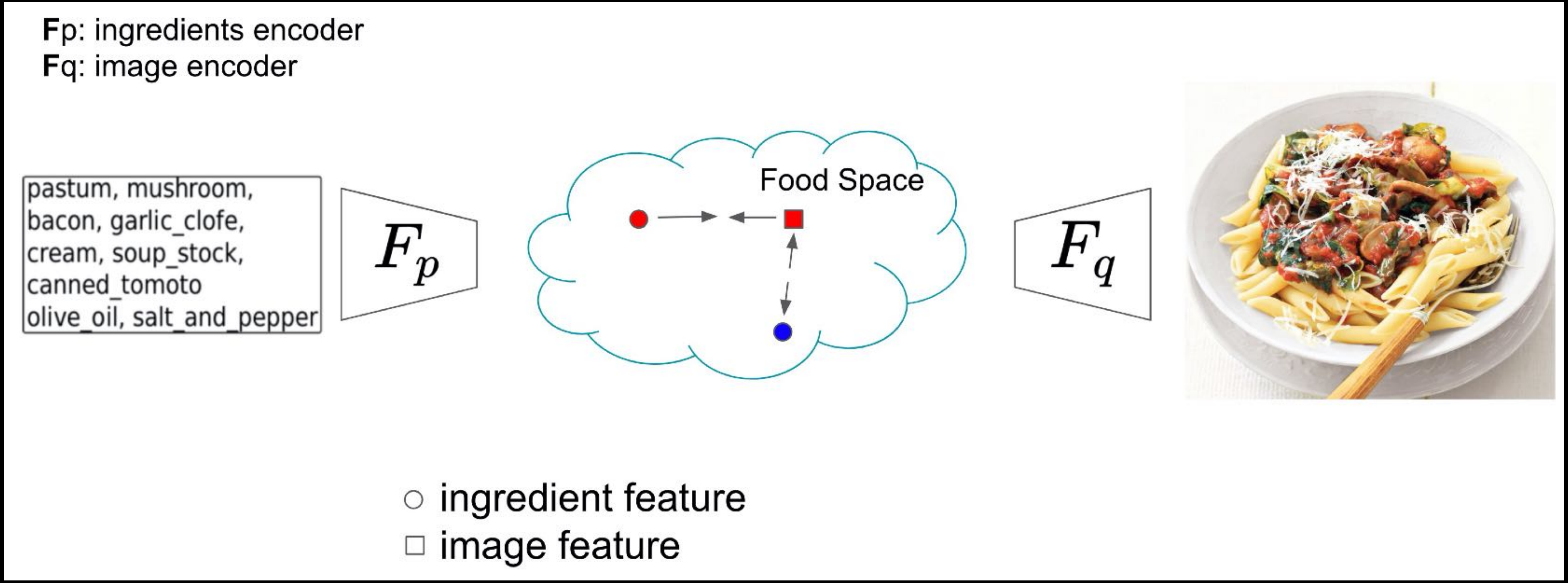
pastum, mushroom,
bacon, garlic_clofe,
cream, soup_stock,
canned_tomoto
olive_oil, salt_and_pepper

Q1 → Q2 →

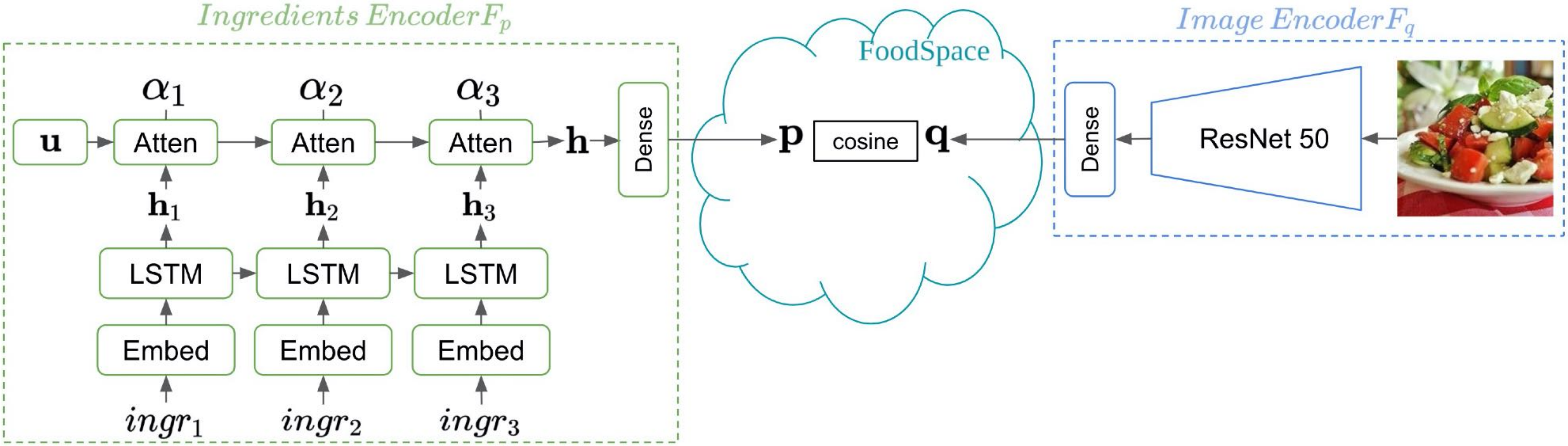
Q1: How to encoder ingredients?

Q2: How to generate image?

Q1 => Attention-based Cross-Modal Association Model

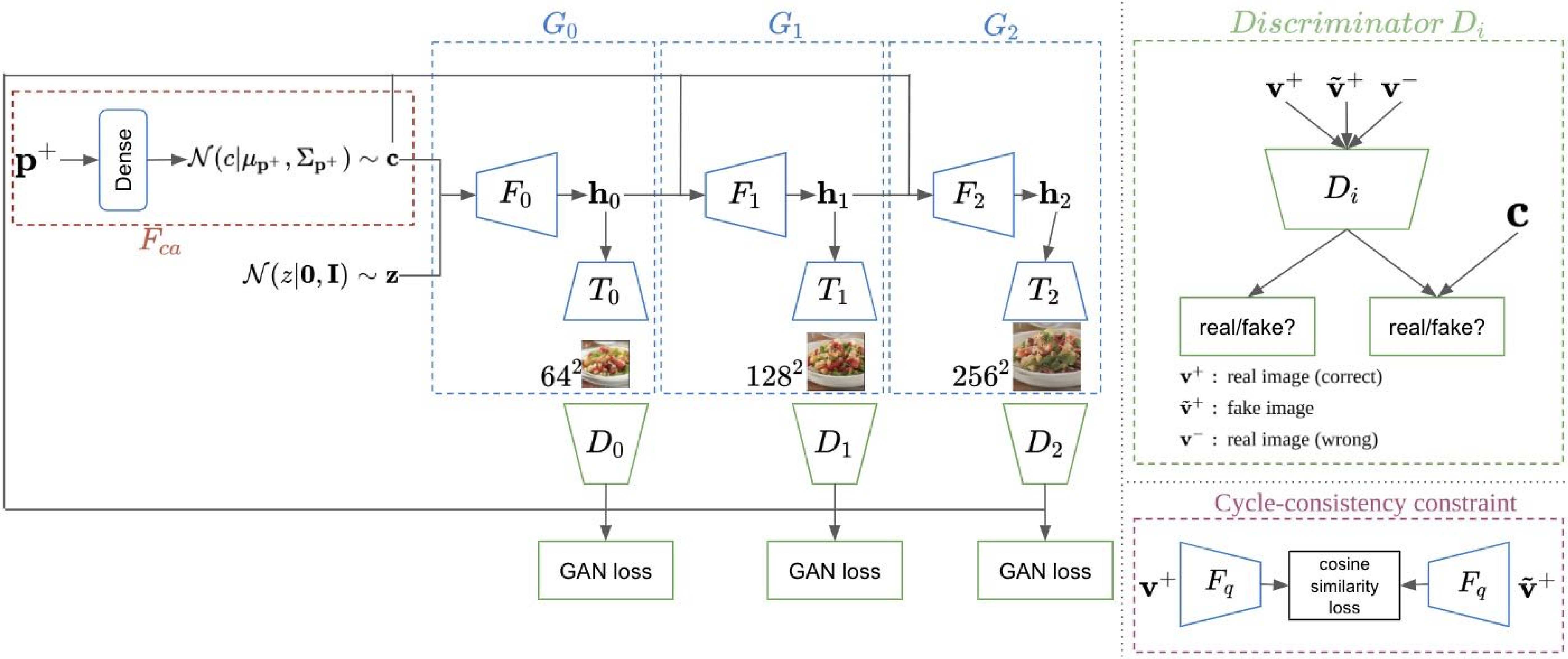


Q1 => Attention-based Cross-Modal Association Model

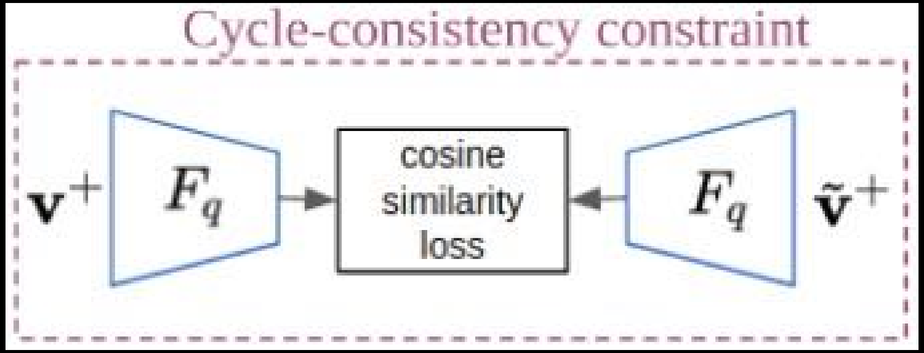


$$V(\mathrm{F_p}, \mathrm{F_q}) = \mathbb{E}_{\hat{p}(\boldsymbol{r}^+, \boldsymbol{v}^+), \hat{p}(\boldsymbol{v}^-)} \min\left(\left[\cos\left[\boldsymbol{p}^+, \boldsymbol{q}^+\right] - \cos\left[\boldsymbol{p}^+, \boldsymbol{q}^-\right] - \varepsilon\right], 0\right) + \mathbb{E}_{\hat{p}(\boldsymbol{r}^+, \boldsymbol{v}^+), \hat{p}(\boldsymbol{r}^-)} \min\left(\left[\cos\left[\boldsymbol{p}^+, \boldsymbol{q}^+\right] - \cos\left[\boldsymbol{p}^-, \boldsymbol{q}^+\right] - \varepsilon\right], 0\right),$$

Q2 => Generative Meal Image Network



Cycle-consistency Constraint

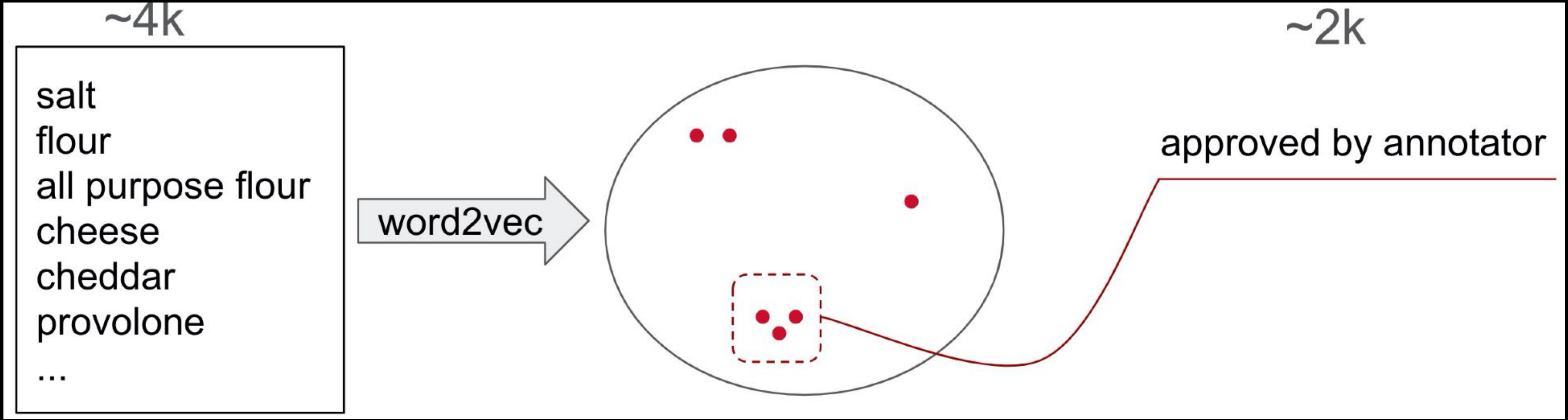


$$\mathcal{L}_i^{cycle} = \cos\left[\boldsymbol{q}^+, \tilde{\boldsymbol{q}}^+\right]$$

$$\mathcal{L}_G = \sum_{i=0}^{2}\left\{\mathcal{L}_i^{cond} + \lambda_{uncond}\,\mathcal{L}_i^{uncond} + \lambda_{cycle}\,\mathcal{L}_i^{cycle}\right\} - \lambda_{ca}\,\mathcal{L}_{ca}$$

Dataset

- Recipe1M dataset [*]
- ~400k recipes with title, **ingredients**, instructions and **images**.
- ~16k ingredients names
  - => reduced to ~4k by frequency
  - => further merged to ~2k by semi-automatic fusing process



[*] Marin, Javier, et al. "Recipe1M: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images." arXiv preprint arXiv:1810.06553 (2018).

Evaluate attention-based association model

		im2recipe				recipe2im			
		MedR↓	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑
1K	attention [5]	-	-	-	-	-	-	-	-
	ours	5.500	0.234	0.503	0.618	5.750	0.230	0.491	0.615
5K	attention [5]	71.000	0.045	0.135	0.202	70.100	0.042	0.133	0.202
	ours	**24.000**	**0.099**	**0.265**	**0.364**	**25.100**	**0.097**	**0.259**	**0.357**
10K	attention [5]	-	-	-	-	-	-	-	-
	ours	47.700	0.065	0.185	0.267	48.300	0.061	0.178	0.261

Tab. 1: Comparison with attention-based association model for using image as query to retrieve recipe. '↓' means the lower the better, '↑' means the higher the better, '-' stands for score not reported in [5].

** 1K, 5K, 10K means how many recipes/images we're retrieving from, the larger the more difficult*

[5]: Chen, Jing-Jing, et al. "Deep Understanding of Cooking Procedure for Cross-modal Recipe Retrieval."

Evaluate generative meal image network

train/test dataset are 17209/3784 (salad), 9546/2063 (cookie) and 4312/900 (muffin)

		salad	cookie	muffin
IS ↑	StackGAN-v2	3.07	**4.70**	2.60
	ours	**3.46**	2.82	**2.94**
	real	5.12	5.70	4.20
FID ↓	StackGAN-v2	**55.43**	106.14	104.73
	ours	78.79	**87.14**	**81.13**

- Training time: 3~4 days for 300 epochs
- Inception Score (IS): higher score means images that are both meaningful and diverse
- Frechet Inception Distance (FID): Frechet distance between real and synthesized data distributions in feature space

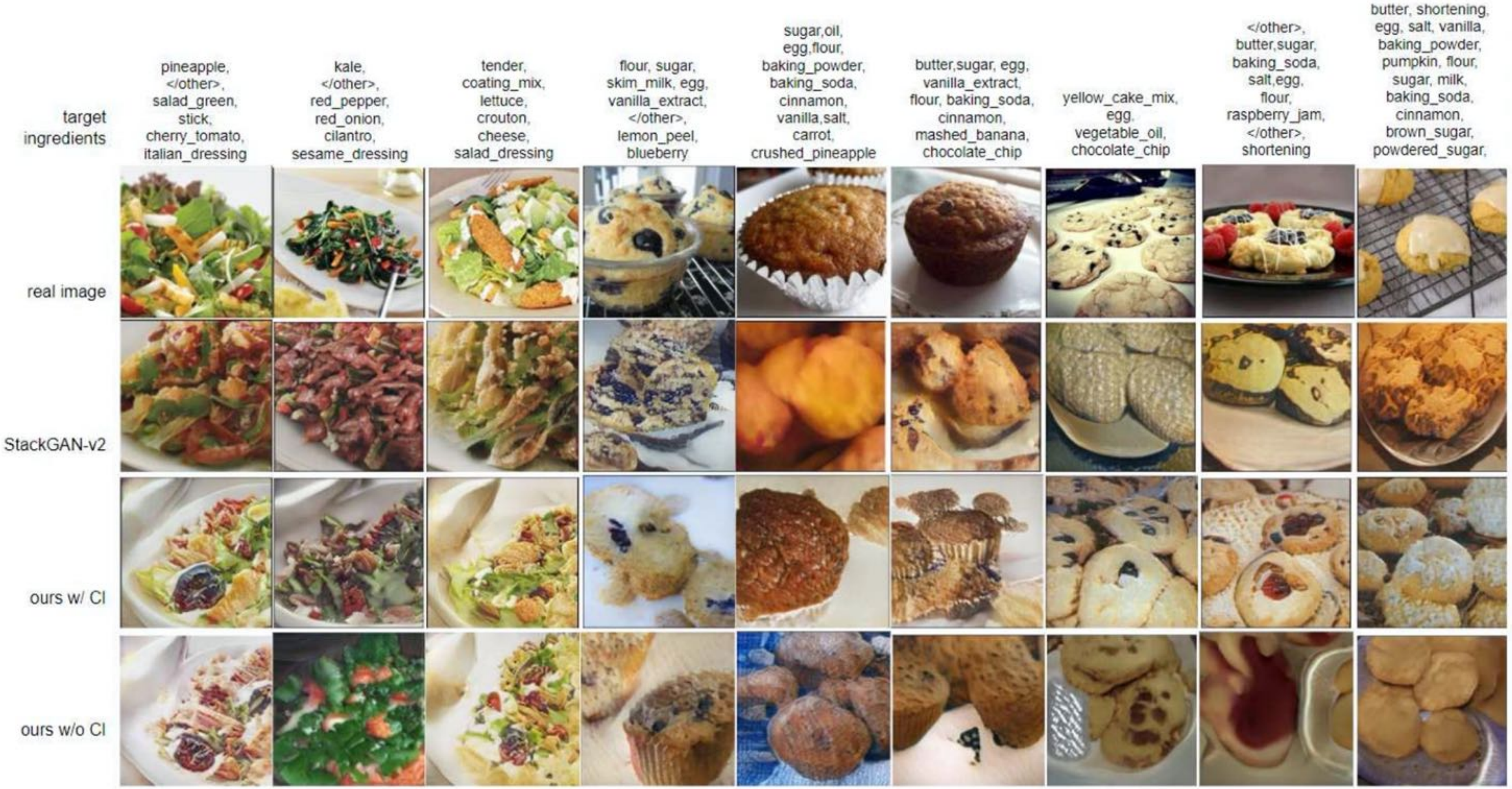
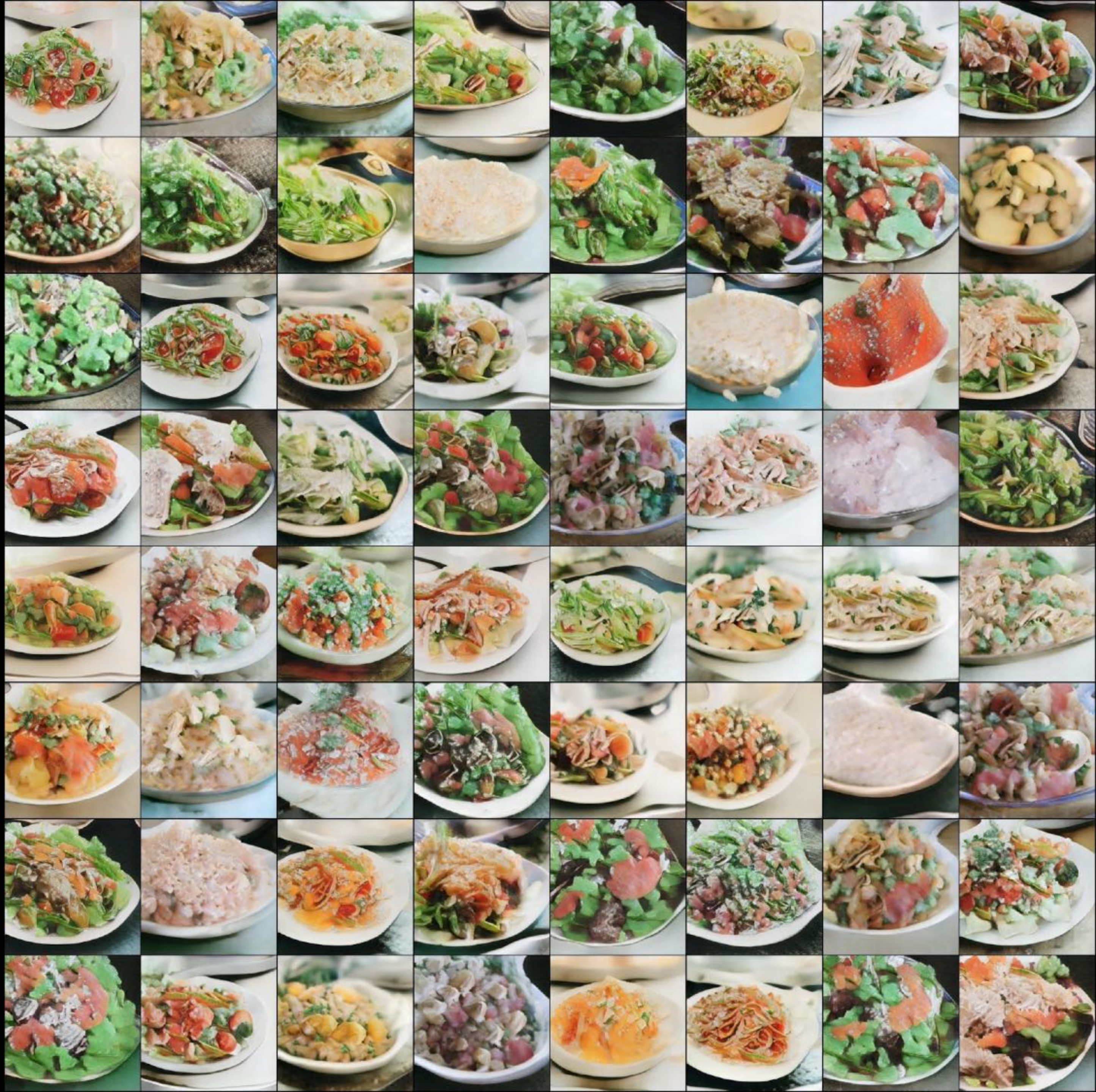


Figure 5: Example results by StackGAN-v2 [27] and our model conditioned on target ingredients, the real images are also shown for reference.

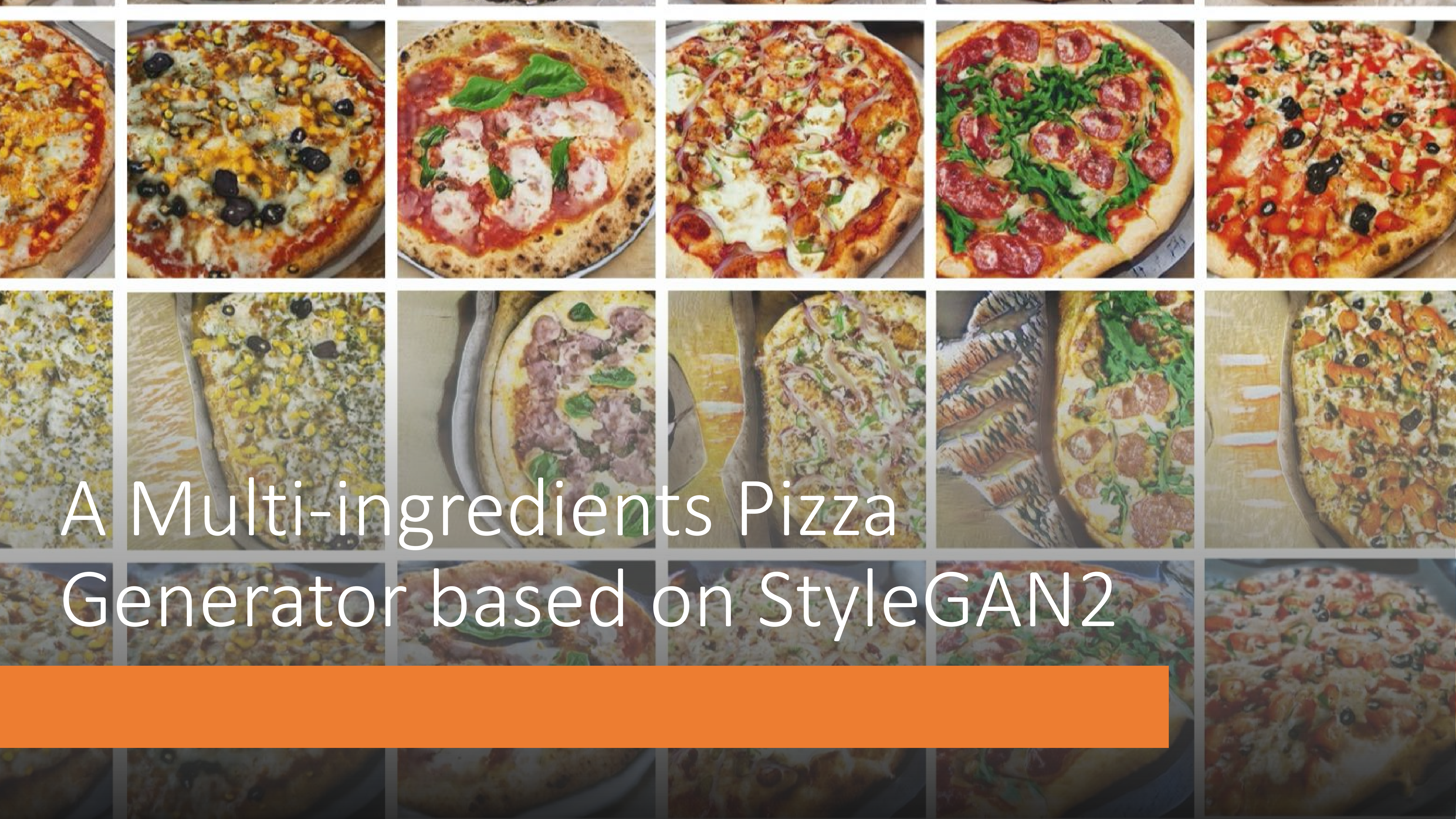
Change Ingredients;
Fix Style Vector



Fix Ingredients;
Change Style Vector

Fig. 5: Example results from same ingredients with different random vectors. 16 synthesized images are shown for each real image (top-left).

Demo
on
foodai.cs.rutgers.edu



A Multi-ingredients Pizza Generator based on StyleGAN2

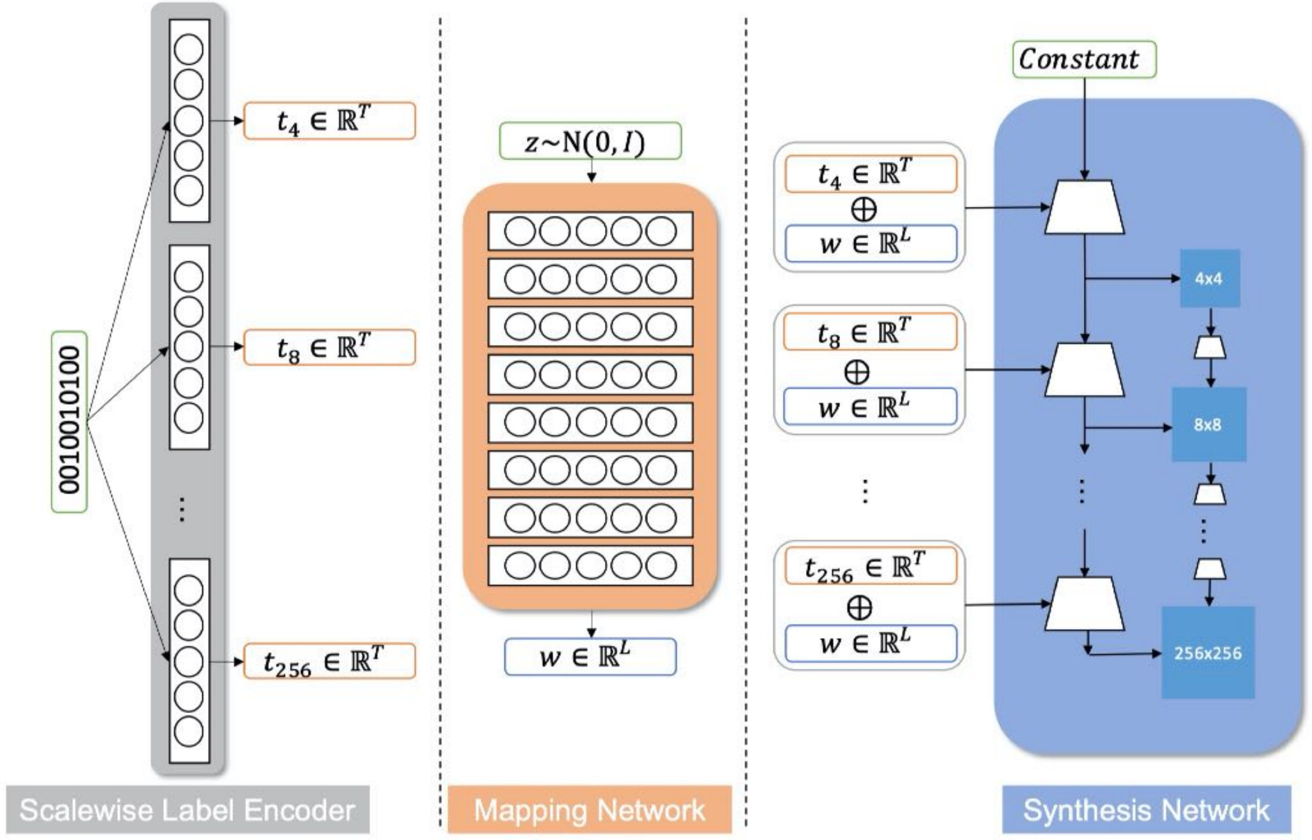


Figure 1. Generator components overview. **Left**: Scalewise Label Encoder (*SLE*), each scale has its own layers for encoding. **Middle**: Mapping Network. **Right**: Synthesis Network, $\oplus$ means concatenation, notice images at different scales are conditioned with different label embedding

Losses

$$\max_G = D(\{\mathbf{t}_i\}, G(\mathbf{x}, \mathbf{z})) + \lambda_c D(G(\mathbf{x}, \mathbf{z})) + \lambda_{clf} BCE\left(\mathbf{x}, h(\tilde{\mathbf{y}})\right),$$

$$\begin{aligned}\min_D = &D(\{\mathbf{t}_i\}, G(\mathbf{x}, \mathbf{z})) + \lambda_{uncond} D(G(\mathbf{x}, \mathbf{z})) \\ &- D(\{\mathbf{t}_i\}, \mathbf{y}) - \lambda_{uncond} D(\mathbf{y}) \\ &+ \lambda_{match} D(\{\mathbf{t}_i\}, \bar{\mathbf{y}}) \\ &+ \lambda_{r1} \left(\frac{\partial D(\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial D(\mathbf{x}, \{\mathbf{t}_i\})}{\partial \mathbf{x}} \right).\end{aligned}$$

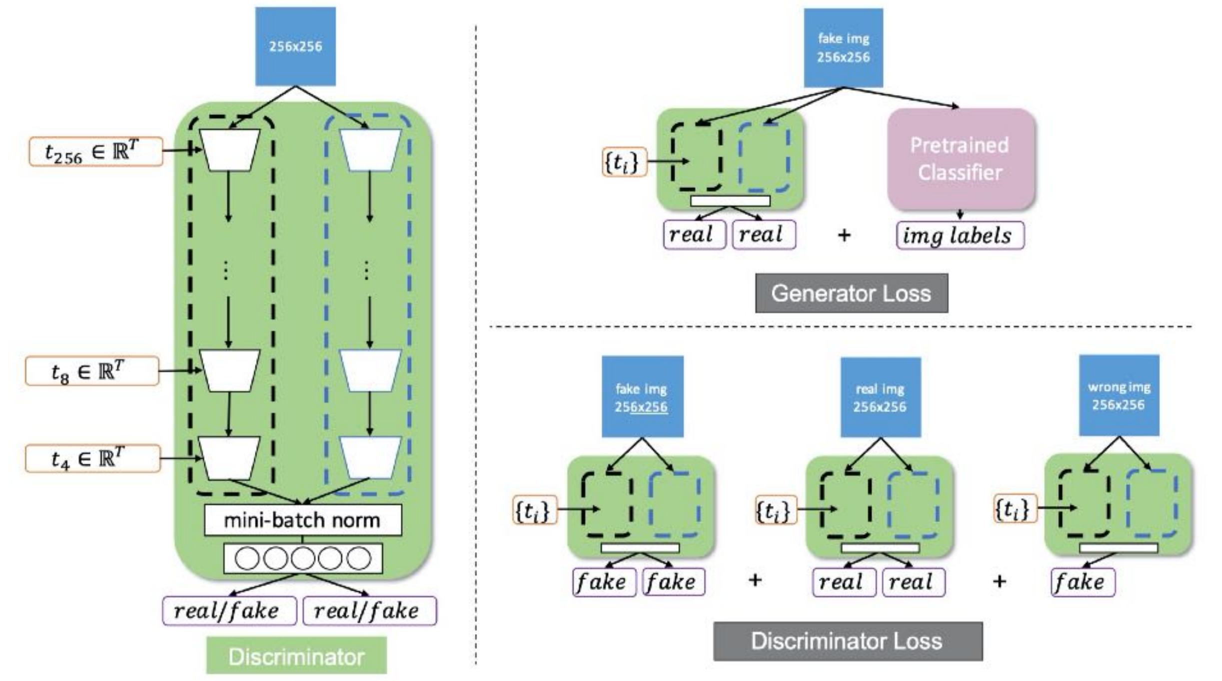


Figure 2. **Left**: Discriminator structure, which contains one branch for conditional output and another branch for unconditional output, notice the label embedding $\{\mathbf{t}_i\}$ are reversed to match different scales. **Top right**: Generator Loss, consisting discriminator loss plus Classification Regularizer (*CR*) loss for fake image. **Bottom right**: Discriminator Loss, consisting three discriminator losses from fake, real and wrong images. Notice the discriminator is trained to distinguish between (txt, real img) and (txt, wrong img)

Table 2. Quantitative comparison of performances between baselines and the proposed multi-ingredient Pizza Generator (*MPG*)

Models	Image Size	FID↓	mAP↑
StackGAN2 [41]	256^2	81.01	0.2001
CookGAN [10]	256^2	81.86	0.3084
AttnGAN [40]	256^2	74.47	0.5729
MPG	256^2	**22.54**	**0.9946**

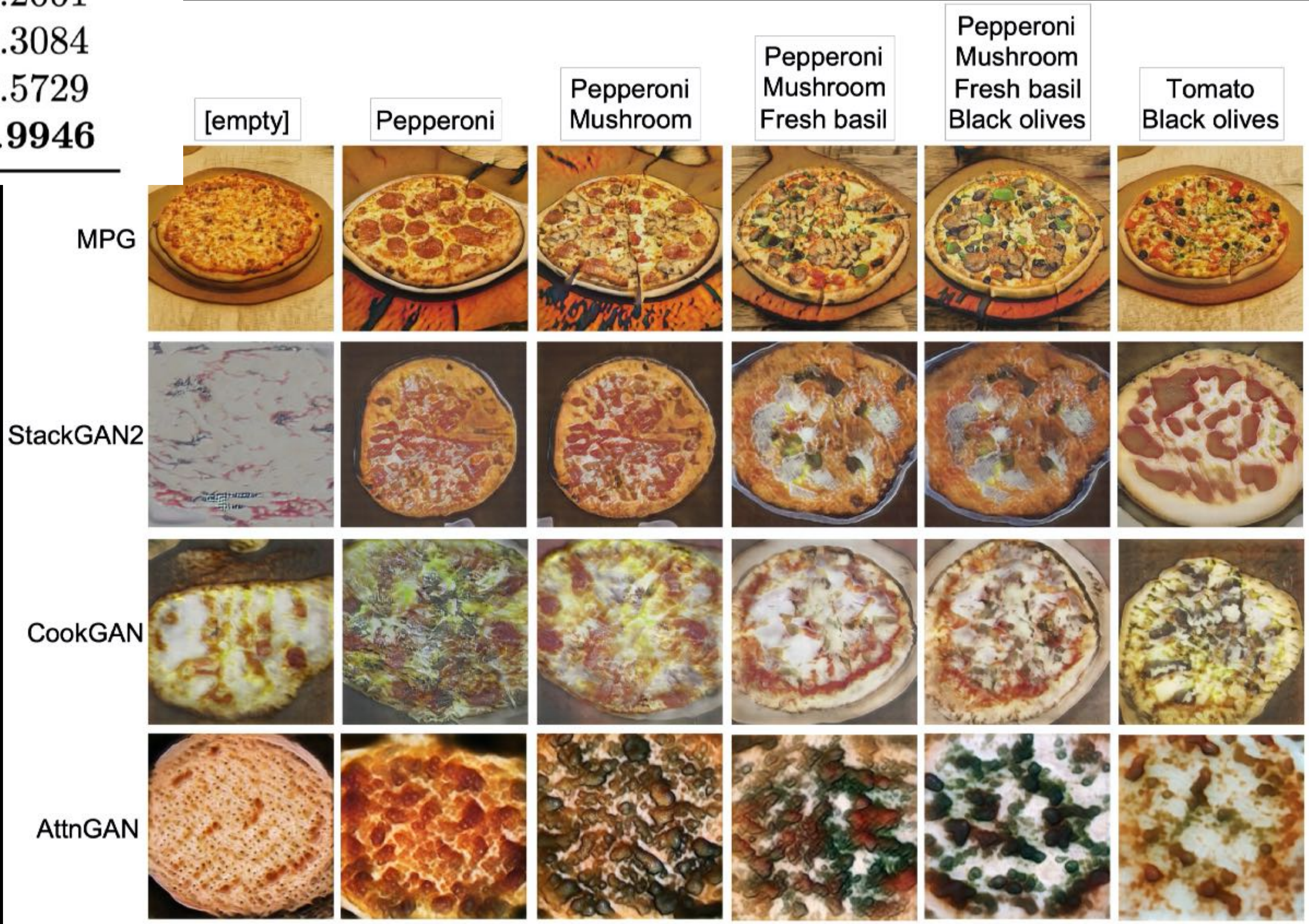


Figure 7. Qualitative comparison between our model *MPG* and baselines

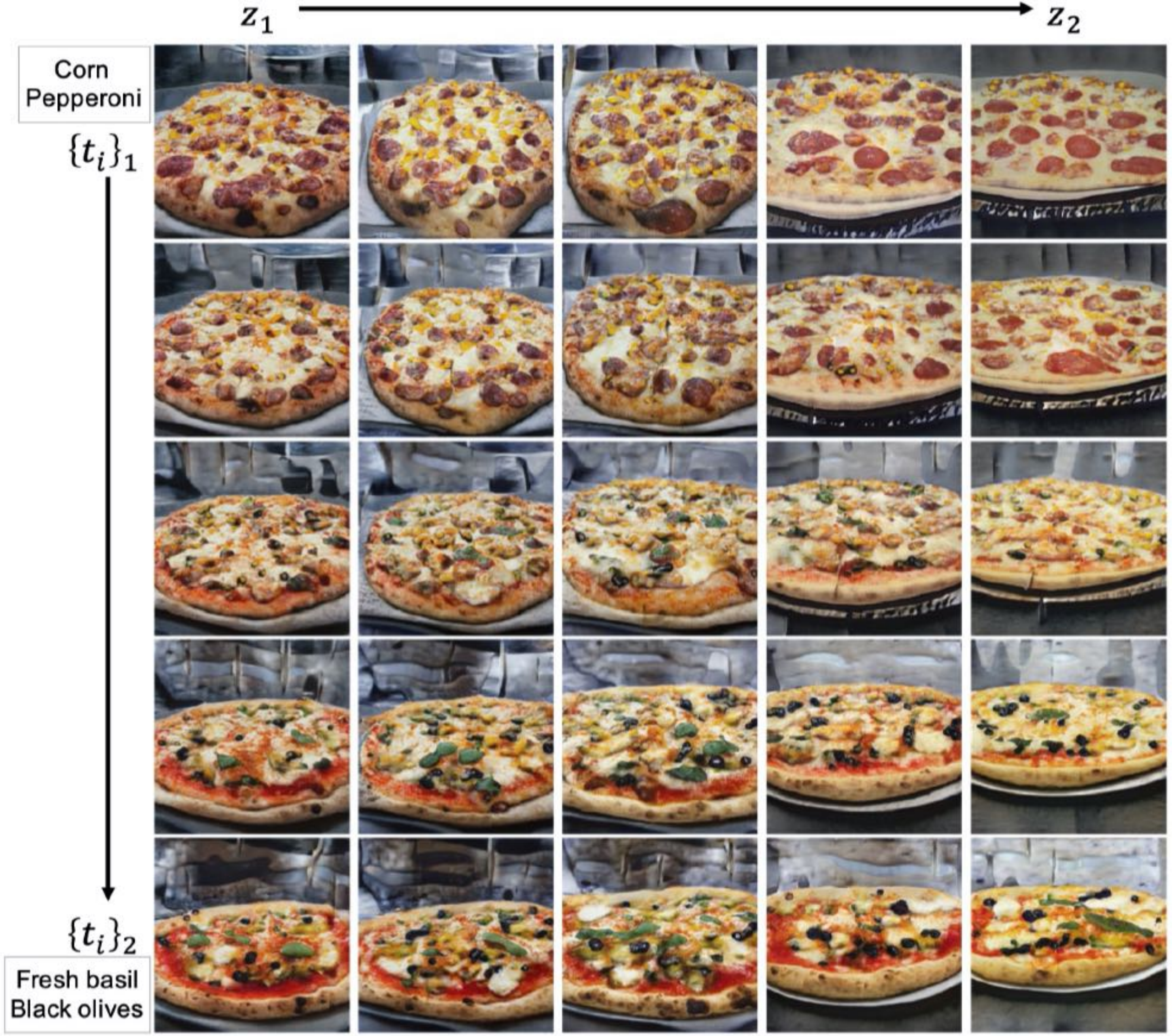


Figure 9. Illustration of traversing through the text embedding space and the style noise space. Images in each row are generated with the same text embedding (interpolated between $\{\mathbf{t}_i\}_1$ and $\{\mathbf{t}_i\}_2$), while images in each column are synthesized with the same style noise (interpolated between $\mathbf{z}_1$ and $\mathbf{z}_2$).

Demo
on
foodai.cs.rutgers.edu

Q & A

- Han, Fangda, Ricardo Guerrero, and Vladimir Pavlovic. "CookGAN: Meal Image Synthesis from Ingredients." *The IEEE Winter Conference on Applications of Computer Vision*. 2020.
- Han, Fangda, et al. "MPG: A Multi-ingredient Pizza Image Generator with Conditional StyleGANs." *arXiv preprint arXiv:2012.02821* (2020).