

MIRecipe: A Recipe Dataset for Stage-Aware Recognition of Changes in Appearance of Ingredients

Yixin Zhang¹, Yoko Yamakata² and Keishi Tajima¹

¹Kyoto University ²The University of Tokyo

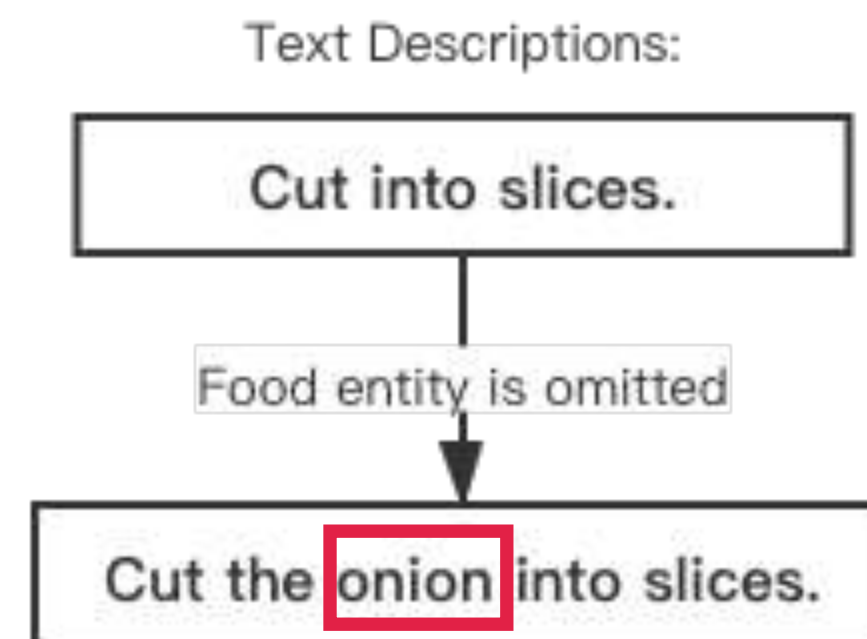
Research Background

- In recent years, user-submitted recipe sites have become popular.
- Recipes with one-to-one correspondence between images and texts
 - rich and valuable multimedia information.
- Recipe website: Haodou
 - About 400,000 recipes
 - Every instructional step is **associated with an image.**



Research Problem

- **Ingredients are omitted** in some instructional steps.
 - This can make the recipe or a particular procedure difficult to understand.
 - For example, when people interacting with smart speakers.
 - However, those entities omitted in text descriptions are sometimes shown in the **attached images**.



Extract ingredient information from images

- Therefore, we need to **recognize ingredient** in instructional images.

Difficulty of the Problem

- Basic image recognition method:
 - Good for entities like tools: shapes of tools are stable



Beginning Stage



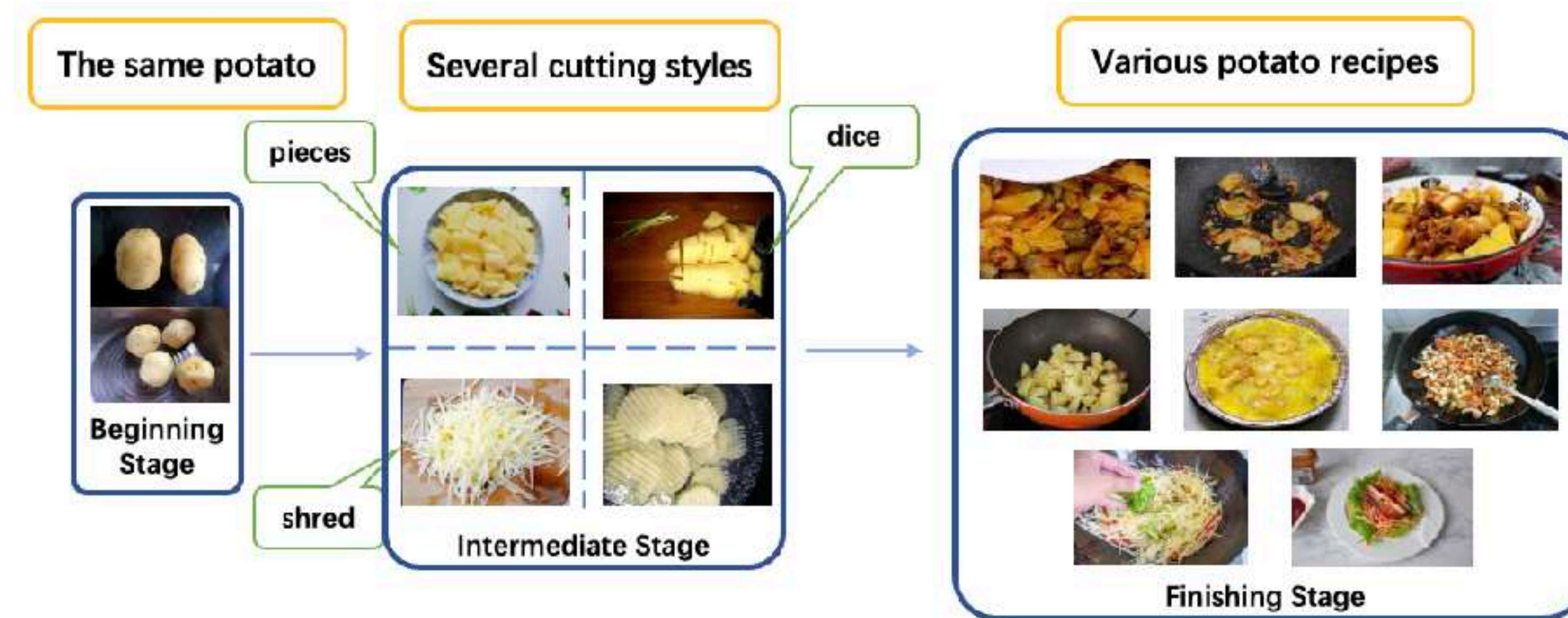
Intermediate Stage



Finishing Stage

Difficulty of the Problem

- Basic image recognition method:
 - Good for entities like tools → applying to tool recognition
 - Difficult for ingredient (ingredient shapes are changing as being cooked) → need adjusting
- Images located in different positions in the cooking procedure of the same ingredient may be very different.



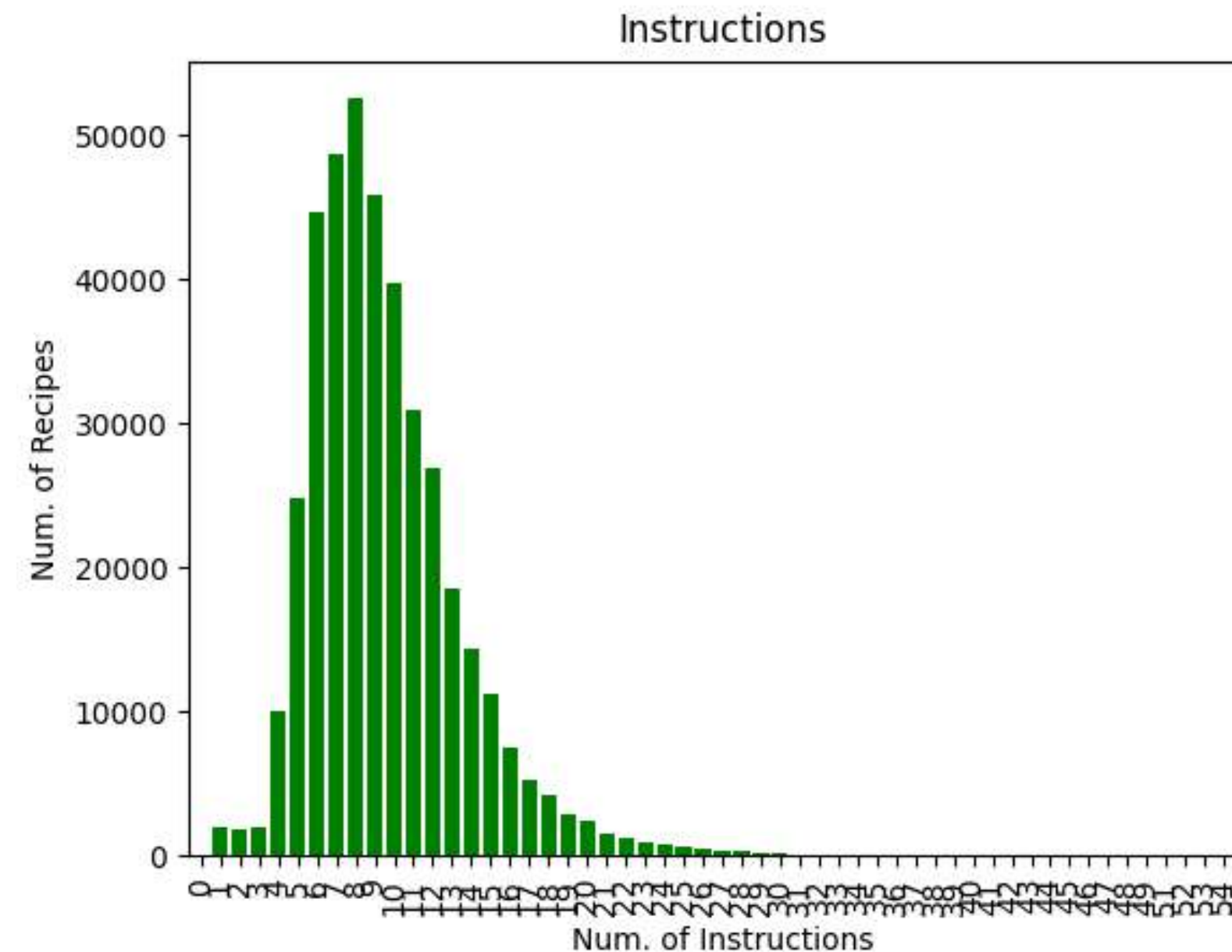
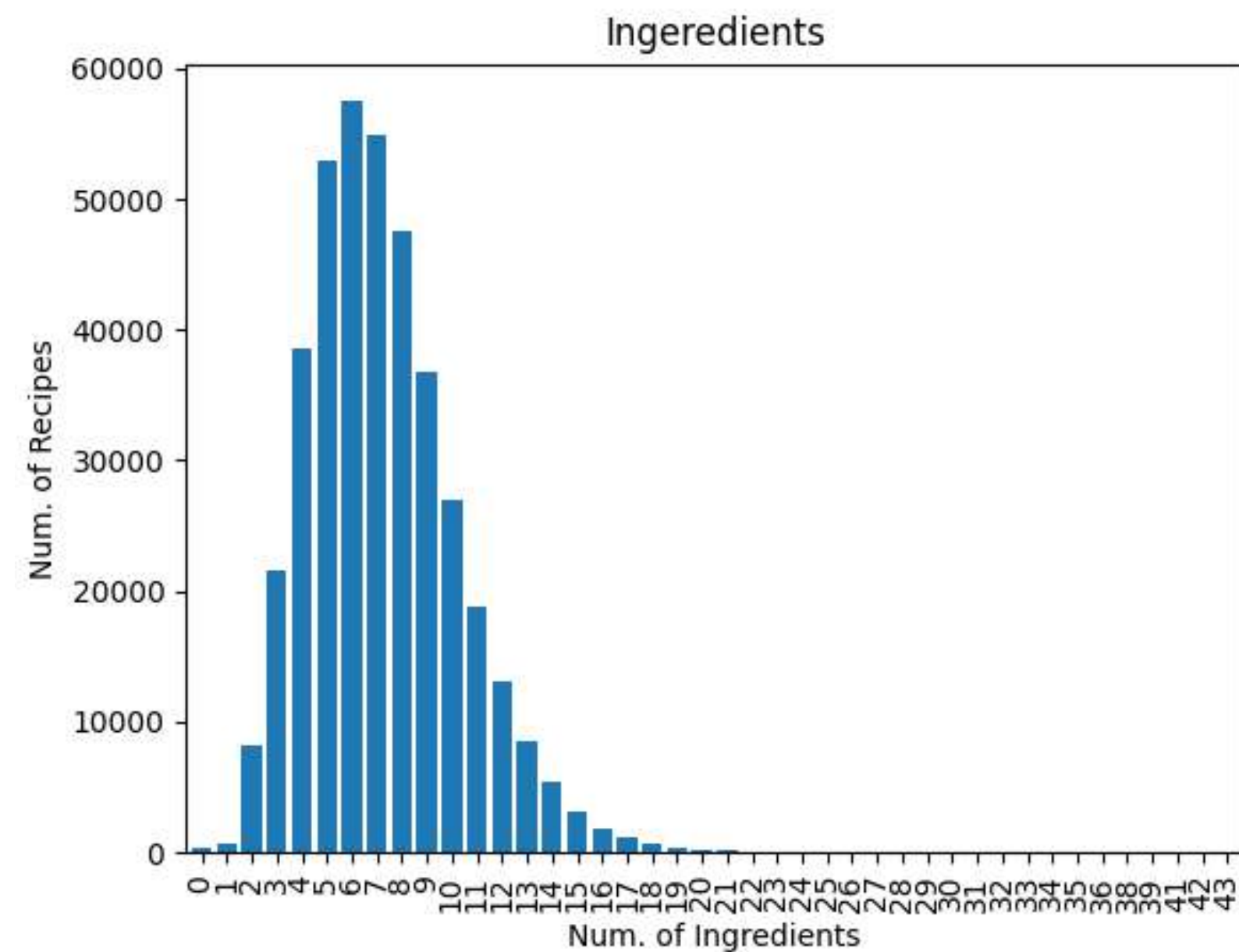
- Thus, we would like to use the features of relative position of steps in the whole recipe.

Research Purpose

- The existing image recognition method barely takes the problem of objects **changing in appearance** into account.
- **Main works**
 - Construct a text-image instructional recipe dataset: **MIRecipe**.
 - Stage-aware image recognition method
 - Recognize **appearance changing ingredients** in recipe instructional images.
 - In order to improve the recognition accuracy.

Dataset

- Recipe Dataset Statistics (up to now)
 - Recipe Num. : 398,597
 - Ingredient Class Num. : 35,319
 - Instruction Num.(text and image): 3,745,544



Dataset

- #Ingredient classes used in this experiments: 20
- high frequencies of occurrence

Potato, ginger, onion, pork, shrimp, chicken, corn, carrot, eggplant, shallot, tofu, spinach, sauce, chili, bread, dough, fish, egg, cucumber, soybean

- #Images of 20 classes: 35,401

Table 9: Division of 20 Classes of Food

	potato	ginger	onion	pork
Subset1	499	1887	330	461
Subset2	409	2108	320	231
Subset3	413	638	175	72
Total	1321	4633	825	764
	shrimp	chicken	corn	carrot
Subset1	730	170	806	856
Subset2	600	155	507	653
Subset3	440	108	246	406
Total	1770	433	1559	1915
	eggplant	shallot	tofu	spinach
Subset1	169	1980	479	221
Subset2	120	2147	342	109
Subset3	108	1725	283	96
Total	397	5852	1104	426
	sauce	chili	bread	dough
Subset1	34	199	618	1574
Subset2	114	184	280	2240
Subset3	220	204	454	1003
Total	378	587	1352	4817
	fish	egg	cucumber	soybean
Subset1	1081	1690	222	129
Subset2	900	1143	187	104
Subset3	824	698	194	96
Total	2805	3531	603	329

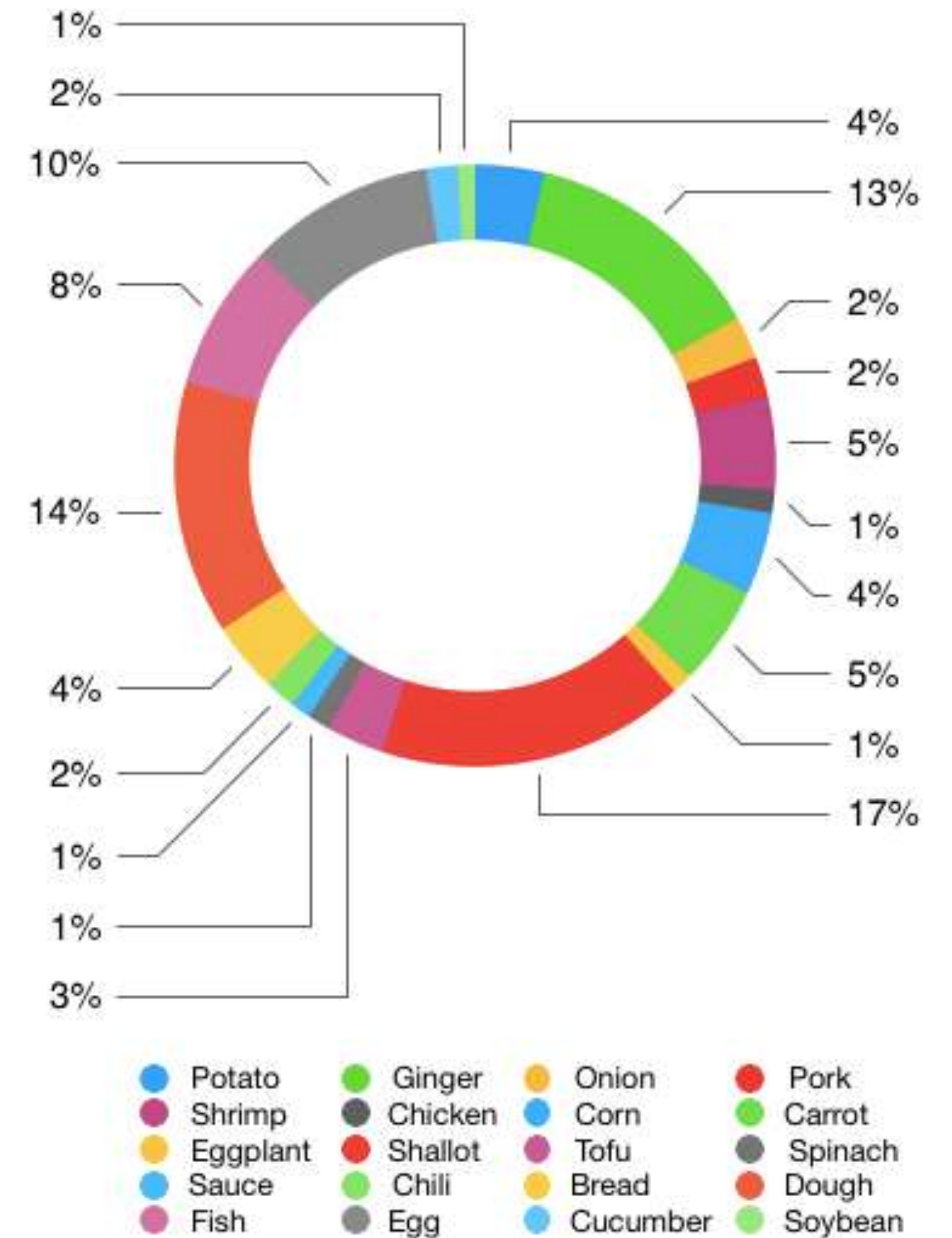


Image Classification

- We compute the relative position of the steps in the whole recipe
 - E.g., Step No.4 out of 15 steps: 0.267

$$RelativePosition = \frac{StepNum.}{TotalStepNum.}$$

- Images are divided into 3 subsets according to their relative positions in recipes.










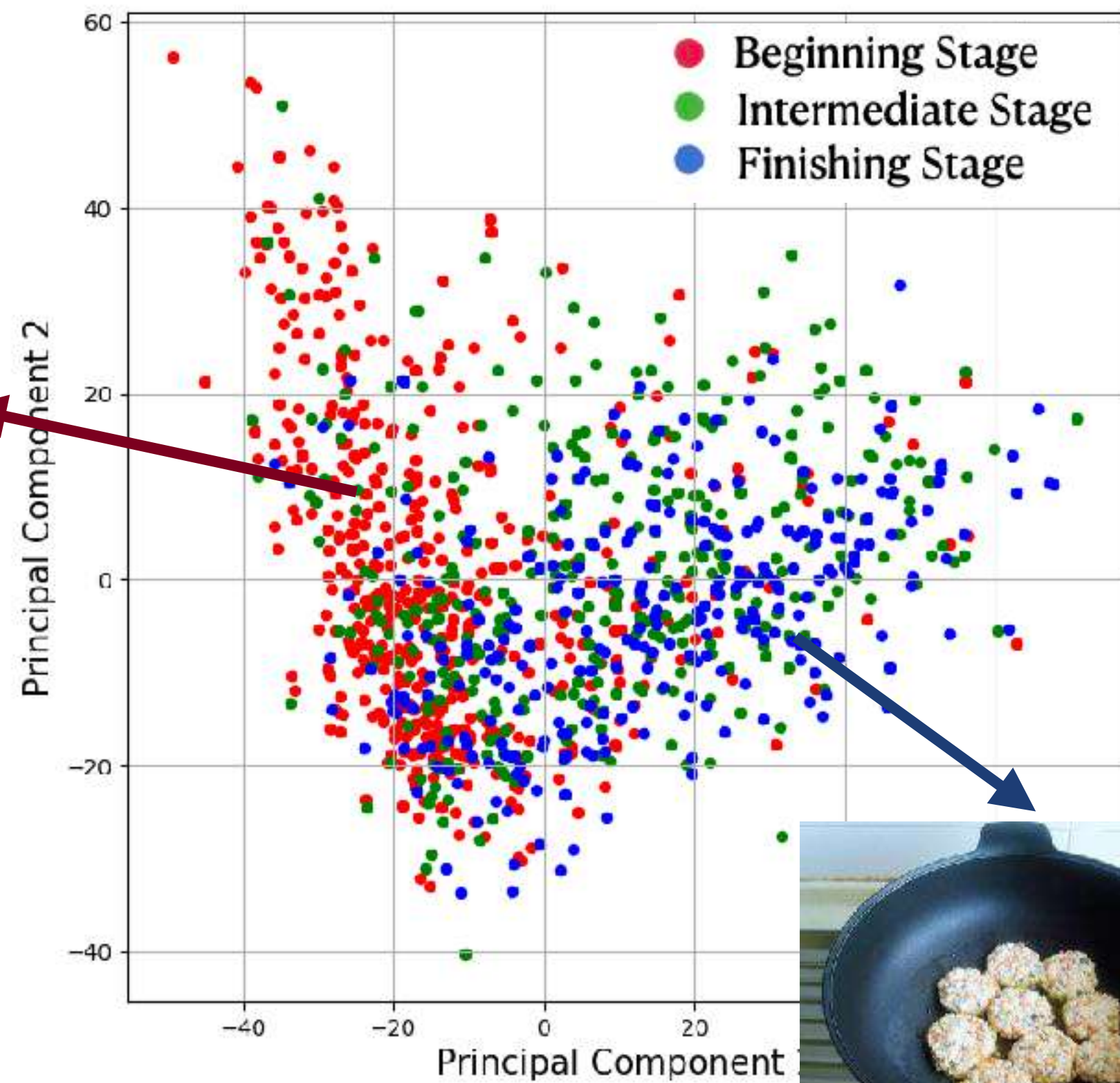
	Recipe A:	Recipe B:	Recipe C:
Subset 1: Beginning Stage	<ul style="list-style-type: none"> Step 1 Step 2 Step 3 Step 4 	<ul style="list-style-type: none"> Step 1 Step 2 Step 3 	<ul style="list-style-type: none"> Step 1 Step 2 Step 3 Step 4 
Subset 2: Intermediate Stage	<ul style="list-style-type: none"> Step 5 Step 6 Step 7 Step 8 	<ul style="list-style-type: none"> Step 4 Step 5 Step 6 	<ul style="list-style-type: none"> Step 5 Step 6 Step 7 Step 8 
Subset 3: Finishing Stage	<ul style="list-style-type: none"> Step 9 Step 10 Step 11 Step 12 	<ul style="list-style-type: none"> Step 7 Step 8 Step 9 	<ul style="list-style-type: none"> Step 9 Step 10 Step 11 Step 12 Step 13 

Image Features Visualization

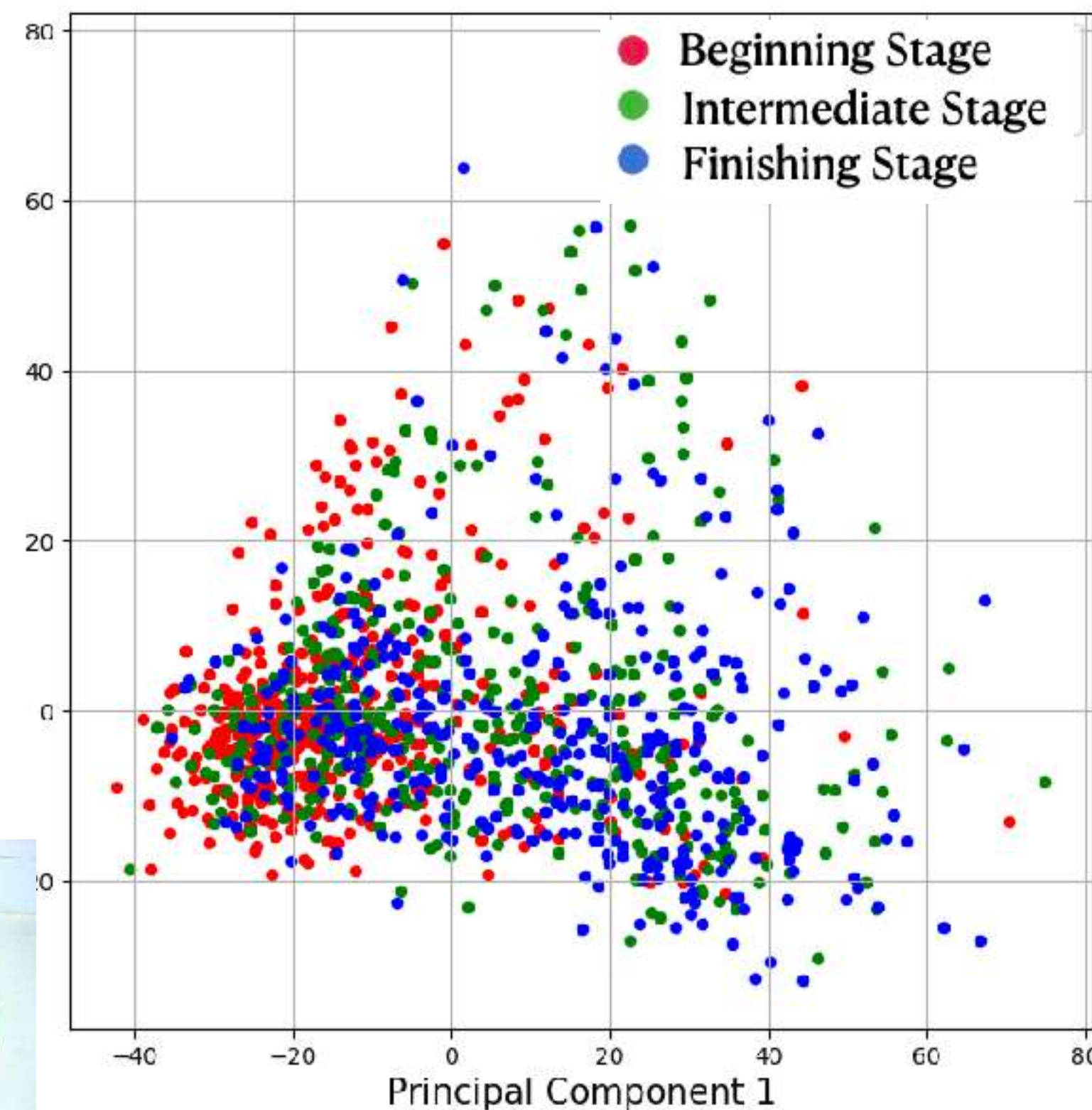
- Visualized embedded feature for the ingredient class per **stage**.
- Images of the same stage have similar characteristics.



Beginning Stage
&
Finishing Stage:
distributed
separately



Tofu

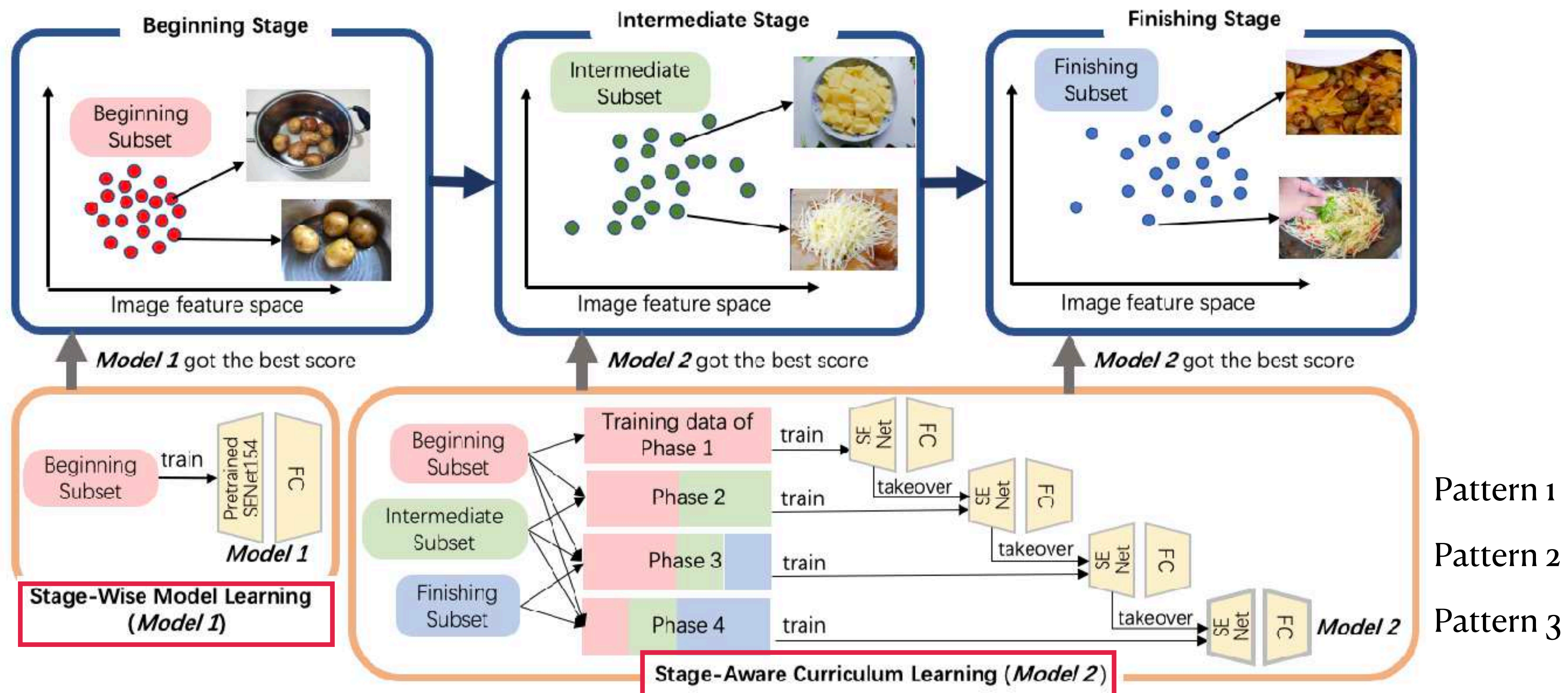


Potato

Beginning
Stage: denser,
Finishing Stage:
scattered

Stage-Aware Ingredient Recognition

- Stage-Aware Recipe Image Recognition For Ingredients Changing in Appearance
- Overview of the baseline method and proposed methods:



Experiment Result

- Comparison of methods

Table 2: Accuracy of Stage-wise Model Learning

Training \ Test Subset	Beginning	Intermediate	Finishing	All
Beginning	64.16%	54.72%	40.63%	49.91%
Intermediate	55.63%	60.59%	49.51%	47.33%
Finishing	42.79%	50.66%	52.83%	47.01%
All	50.17%	51.74%	46.28%	49.63%

Table 3: Accuracy of Stage-aware Curriculum Learning

Training Pattern \ Test Subset	Beginning	Intermediate	Finishing
Pattern 1	64.16%	54.72%	40.63%
Pattern 2	61.79%	58.84%	55.31%
Pattern 3	60.10%	62.61%	58.34%

Table 4: Comparison of the Proposed and Baseline Methods

Plan		Top-1 acc.
Baseline (SENet154)		49.63%
Stage-Wise Model Learning ($m(i) = i$)	Beginning Subset	64.16%
	Intermediate Subset	60.59%
	Finishing Subset	52.83%
	Average	59.19%
Curriculum Learning Pattern 3 (Model 2)	Beginning Subset	60.10%
	Intermediate Subset	62.61%
	Finishing Subset	58.34%
	Average	60.35%

Experiment Result

- Comparison of methods

Table 5: Final Accuracy of Our method

	Model Selection	Accuracy
Beginning Subset	Model 1	64.16%
Intermediate Subset	Model 2	62.61%
Finishing Subset	Model 2	58.34%
Average		61.70%

Table 6: Comparison of Our Methods Based on SENet154 with Other Standard Methods

	Plan	Top-1 acc.	Top-3 acc.	Top-5 acc.
Ours	Stage-Wise	59.19%	81.21%	89.47%
	Model 2	60.35%	83.76%	90.91%
Baseline	SENet154	49.63%	76.51%	86.93%
	Resnet50	46.41%	74.13%	84.35%
	VGG16	43.67%	72.39%	85.01%
	AlexNet	31.77%	64.06%	77.42%

Conclusion

- A recipe dataset
 - **contains both text and image data for every cooking step.**
- A recognition method
 - for **ingredients whose appearance changes** with the cooking progress.
- We only focused on the single-label recognition in this work.
 - Experiments with multi-label data is also an important remaining issue for future work.

Thanks for listening!