# Chenyun Wu - Research Statement

chenyun@cs.umass.edu
https://people.cs.umass.edu/~chenyun

My general research goal is to build artificial intelligence systems that can perceive visual signals, communicate with people in the visual context, and take actions based on its perception and communication input to assist people in various scenarios. For example, a robot can interact and assist a human to navigate in a scene using language, or an automatic system can edit the visual content of images or videos through natural language commands. Thanks to the development of deep neural networks and large-scale datasets, the fields of computer vision, natural language processing, as well as reinforcement learning have achieved great progress in recent years. However, it is challenging to bring them together and benefit realistic applications. I aim to address these challenges and work on **better understanding of the complex visual world, improving the modeling for more detailed and accurate correlation between visual and language signals, and better reasoning based on perceived information.**

My thesis work is about joint modeling of vision and language, with an emphasis on a better understanding of visual domains. I have also worked on reasoning and decision-making based on visual perception. I will elaborate further in both directions.

## Joint Understanding of Vision and Language

Vision and language are the fundamental media of perception and communication respectively, thus the joint understanding of the two is essential for an intelligent system. Natural language descriptions can provide large-scale and detailed supervision to achieve a better understanding of visual domains. We specifically look into four visual domains: (1) images of categories within a fine-grained taxonomy, (2) images of texture which describes local patterns, (3) objects and stuff regions in natural images, and (4) clips in videos (as ongoing work). We demonstrate that by aligning visual representations with language, one can enable several applications such as image retrieval and editing, as well as fine-grained classification with naturally interpretable models.

While the representations vary across domains, we address common challenges when combing vision and language. Firstly, tasks need to be carefully designed such that joint understanding is required and cannot be circumvented by only vision or language or guessing based on statistical bias in datasets. Secondly, vision and language tend to cover different contents. Concepts in visual and language worlds follow long-tail distributions and show ambiguity (lack of information) and redundancy (information not useful for the given task) issues in different aspects, which is possible to overcome by combining visual and language clues. Lastly, vision and language differ in structure. Vision is high-dimensional (2D for images and 3D for videos) with hierarchical semantics from color, texture to objects, and relationships between them. Language descriptions on different visual semantic levels possess different vocabularies. Language is discrete and compositional (e.g., an entity can be modified by various attributes). It is challenging to align the composition and relationships in language to visual signals.

### Attribute phrases on fine-grained categories [1]

We present a framework for learning to describe fine-grained visual differences between instances using *attribute phrases*. Attribute phrases capture distinguishing aspects of an object (e.g., "propeller on the nose" or "door near the wing" for airplanes) in a compositional manner. Instances within a category can be described by a set of these phrases and collectively they span the space of semantic attributes for a category. We collect a large dataset of such phrases by asking annotators to describe several visual differences between a pair of instances within a category. We then learn to describe and ground these phrases to images in the

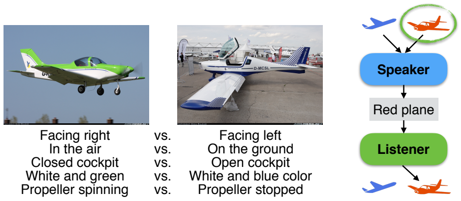
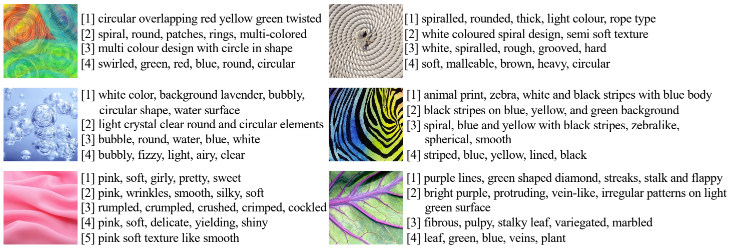

Facing right vs. Facing left
In the air vs. On the ground
Closed cockpit vs. Open cockpit
White and green vs. White and blue color
Propeller spinning vs. Propeller stopped
Speaker
Red plane
Listener

Figure 1: Reference games with attribute phrases.

context of a *reference game* between a speaker and a listener. The goal of a speaker is to describe the attributes of an image that allows the listener to correctly identify it within a pair. The speaker model follows the design of language captioning models while the listener learns to project images and phrases to the same embedding space. We also show that embedding an image into the semantic space of attribute phrases improves fine-grained classification accuracy over existing attribute-based representations. Our speaker model can be applied to describe the main difference between the two given categories. Our listener model is able to retrieve images with a set of attributes.

### Describing textures with natural language [2]

For image classification, especially in fine-grained domains, deep neural networks are known to rely largely on recognizing textures. We focus on natural language for describing textures, which allows us to exclude the effects of shape, object category, or other high-level cues. Textures in natural images can be characterized by color, pattern, periodicity of elements within them, and other attributes that can be described using natural language. We propose a novel dataset containing rich descriptions of textures and conduct a systematic study of current generative and discriminative models for grounding language to images on this dataset. We find that while these models capture some properties of texture, they fail to capture several compositional properties (e.g., "colors of dots"). Our dataset also allows us to train interpretable models and generate language-based explanations of what discriminative features are learned by deep networks for fine-grained categorization where texture plays a key role. We present visualizations of several fine-grained domains and show that texture attributes learned on our dataset offer improvements over expert-designed attributes.



[1] circular overlapping red yellow green twisted
[2] spiral, round, patches, rings, multi-colored
[3] multi colour design with circle in shape
[4] swirled, green, red, blue, round, circular

[1] spiralled, rounded, thick, light colour, rope type
[2] white coloured spiral design, semi soft texture
[3] white, spiralled, rough, grooved, hard
[4] soft, malleable, brown, heavy, circular

[1] white color, background lavender, bubbly, circular shape, water surface
[2] light crystal clear round and circular elements
[3] bubble, round, water, blue, white
[4] bubbly, fizzy, light, airy, clear

[1] animal print, zebra, white and black stripes with blue body
[2] black stripes on blue, yellow, and green background
[3] spiral, blue and yellow with black stripes, zebralike, spherical, smooth
[4] striped, blue, yellow, lined, black

[1] pink, soft, girly, pretty, sweet
[2] pink, wrinkles, smooth, silky, soft
[3] rumpled, crumpled, crushed, crimped, cockled
[4] pink, soft, delicate, yielding, shiny
[5] pink soft texture like smooth

[1] purple lines, green shaped diamond, streaks, stalk and flappy
[2] bright purple, protruding, vein-like, irregular patterns on light green surface
[3] fibrous, pulpy, stalky leaf, variegated, marbled
[4] leaf, green, blue, veins, plant

Figure 2: Describable Textures in Detail Dataset.

### Language-based image segmentation of objects and stuff regions [3]

We extend the referring task to a more realistic setting: instead of selecting one image out of a pair, we consider segmenting image regions given a natural language phrase. We collect a large-scale dataset by automatically constructing referring phrases from Visual Genome, asking human annotators to draw referred regions, and comparing them to Visual Genome annotations for automatic verification. Phrases in our proposed dataset correspond to multiple regions and describe a large number of object and stuff categories as well as their attributes such as color, shape, parts, and relationships with other entities in the image. Our experiments show that the scale and diversity of concepts in our dataset pose significant challenges to the existing state-of-the-art. We systematically handle the long-tail nature of these concepts through an attention mechanism that leverages predictions on easier concepts to improve accuracy on rare concepts. We also adopt a modular approach to combine category, attribute, and relationship cues that can reason on the image grid and output arbitrary referred regions. Our model outperforms existing approaches by a large margin but also calls for further study on rare categories and small target regions.

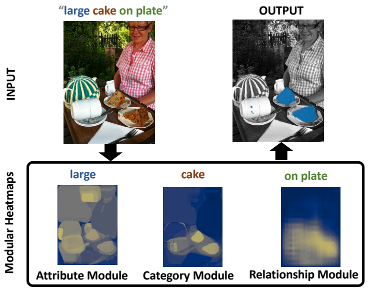


Figure 3: Segmenting image regions given a language phrase.

## Reasoning and Decision Making based on Visual Perception

I have been working on visual geo-localization with action planning as my Master Thesis. The goal is to allow an agent to determine its location in a known environment based on its visual perception. The agent can take actions and move around in the environment to exclude ambiguity of predicted location. Our further goal is to let the agent reach a target location.

We collected both real-world and synthetic data of images and their locations. For real-world data, we adopted the SLAM method to improve ground-truth GPS accuracy for camera data. For synthetic data, we used Minecraft maps that are large-scale and look like cities. We collected data in various scenery, weather, and daylight conditions.

We first train a neural network to predict location from a single image in various conditions. We then implement a particle filter on top to predict locations from a sequence of observations when the agent moves around in the environment. Finally, we design rule-based action planning so that the agent can localize itself more efficiently.

We also tried reinforcement learning methods but didn't outperform rule-based strategies. Possible reasons include: the Minecraft world we selected may not be suitable for the task; the reward signal is extremely sparse; and it requires highly hierarchical planning (e.g., ”turn left at next crossing” requires setting a non-zero speed, stopping when detecting a crossing, rotating the robot to face the correct direction and keep moving forward.)

## Ongoing and Future Directions

### Localizing clips in videos through language descriptions

One of my ongoing projects is to localize clips in videos through language descriptions. Different from referring tasks on images, it requires further temporal reasoning. Most of the existing works encode both the description and video frames, and then either predict on each frame or directly output start and end timestamps. We argue that encoding the whole description into one embedding may lose important information from the structure of the description. We propose an attention mechanism that learns the mapping of each word to video frames, guided by object detection results on each frame. On top of it, we design a graph convolutional network to conduct reasoning based on the syntactic graph of the description. Our preliminary experiments verify the effectiveness of our idea.

There have been more and more explorations of adapting transformer networks on visual tasks. We argue that transformers may be a suitable structure for joint reasoning of language descriptions and video frames. It's on our plan to design and experiment with transformers on our temporal localization task.

### Guided image editing and synthesis

Another of my ongoing project is to edit or generate images based on user guidance. The user input can be in the format of anchor points or sketches on the image, corresponding scales from a selected attribute list, or free-form natural language description. Currently, we are experimenting with texture editing such as changing the blue dots to green with language input. We can push further to edit specific regions in an image (e.g. change the material of the sofa/skirt) and to generate realistic images from simple sketches plus language descriptions.

### Neural network interpretation and analysis

In our work of describing textures, we adopt our collected attribute phrases and metric learning models to generate language interpretations for fine-grained classification models, unveiling what texture features the classifier is looking for when recognizing each category. In our work of language-based segmentation, we can analyze modular output to explain failure cases whether it fails on detecting the category or understanding the attribute or relationship.

I hope to explore further in this direction to associate specific image regions as well as network layers and channels with words or phrases, and to provide interpretation for more types of networks beyond classifiers such as generative adversarial networks, image captioning, and visual question answering models.

### Detection and segmentation of language-based categories

In our language-based segmentation work, we find it challenging to detect and segment categories parsed from language descriptions because it introduces a massive number of categories that follow an extreme long-tail distribution, and the categories are not mutually exclusive. We handle the referring task without fully solving detection by an attention mechanism to leverage detections on easier categories to improve performance on rare categories. However, the detection and segmentation itself is interesting and worth solving. We hope to explore if there are better ways to leverage the correlations and hierarchical structures in language concepts to improve detection and segmentation especially for rare and even unseen categories.

### Dialog based on visual context

While there has been great progress on chatbots in recent years, I think one of the main differences between chatbots and human conversations is that our current systems lack the understanding of context. The visual environment is an important component of the conversation context and I find it an interesting problem to study how the visual context can improve artificial dialogues. For example, the conversation will be very different in a conference room compared to in a household kitchen.

It is related to image captioning and visual question answering tasks but there is a fundamental difference: captioning and VQA tends to focus on information conveyed from the visual signal (such as the number and color of bananas on the table), while dialogues in visual context treat these direct information as shared background knowledge and discuss further questions (such as where and where did you buy the bananas).

### Visual language navigation

Visual language navigation requires a robot to understand natural language instructions, reason them with visual perception, and navigate in a visual environment. It is a highly challenging task that requires the combination of computer vision, natural language understanding, and reasoning. My earlier years of research on geo-localization with action planning is closely related but didn't reach a positive result. In recent years there have been several well-constructed simulators and datasets for this task and a lot of advances in related methodologies. I'm very interested in revisiting this direction.

### Other domains

It would be great if I could continue working on vision and language in the directions discussed above. On the other hand, I also hope to keep open-minded for broad topics in computer vision as long as my previous experience can be useful. I'm especially curious about video understanding and reinforcement learning.

## References

[1] Jong-Chyi Su, Chenyun Wu, Huaizu Jiang, and Subhransu Maji. Reasoning about fine-grained attribute phrases using reference games. In *International Conference on Computer Vision (ICCV)*, 2017.

[2] Chenyun Wu, Mikayla Timm, and Subhransu Maji. Describing textures using natural language. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020.

[3] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020.