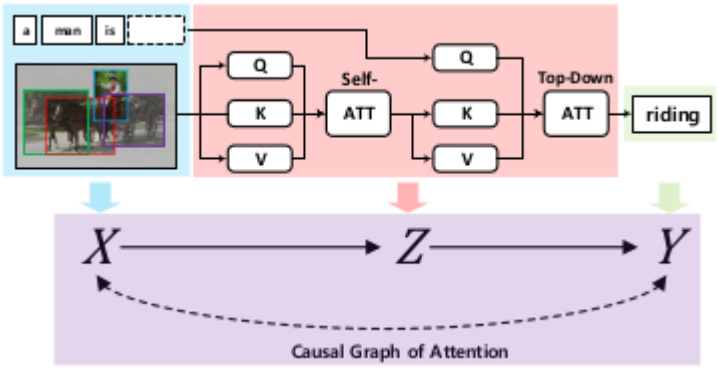
- Causal Attention for Vision-Language Tasks
  - ▶ Maheep's Notes

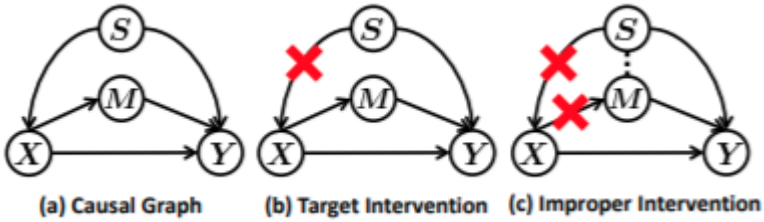
    `P(Z = z|X)` known as **In-Sampling** and a predictor which exploits Z to predict Y.
    `P(Y|X) = sigma P(Z = z|X)P(Y|Z = z)`
    But the predictor may learn the spurious correlation brought by the backdoor path from X to Z, and thus the backdoor method is used to block the path from X to Z, making it:
    `P(Y|do(Z)) = sigma P(X = x)P(Y|X = x,Z)`
    where `P(X = x)` is known as **Cross-Sampling** and making the whole equation:
    `P(Y|do(X)) = sigma P(Z = z|X) sigma P(X = x)P(Y|Z = z, X = x)`



- Causal Attention for Unbiased Visual Recognition
  - ▶ Maheep's Notes

    `M` are retained while the non-causal features `S` are eradicated as shown in the figure below.
    Therefore to disentangle the the `S` and `M`, the equation can be derived as:
    `P(Y|do(X)) = sigma_for_s sigma_for_m P(Y|X, s, m)P(m|X,s)P(s)`

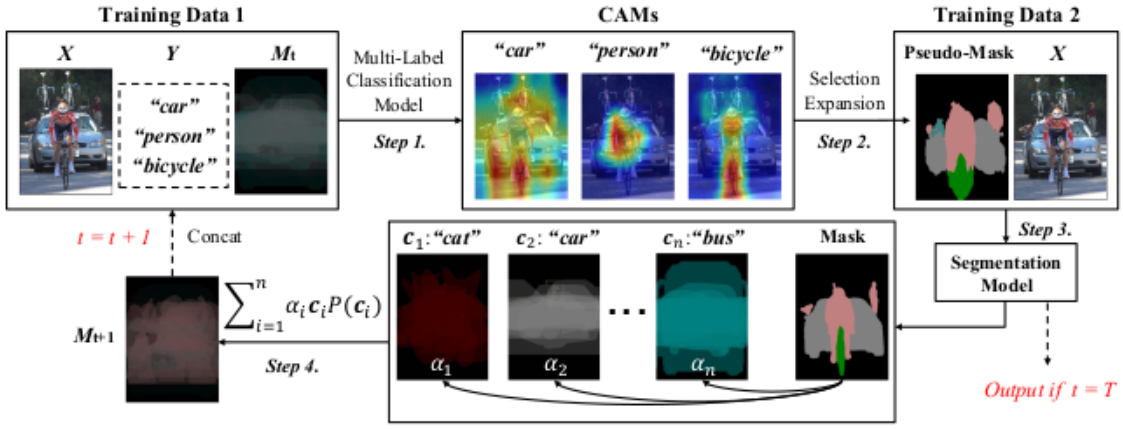
    `P(Z = z|X)` known as **In-Sampling** and a predictor which exploits Z to predict Y.
    `P(Y|X) = sigma P(Z = z|X)P(Y|Z = z)`
    But the predictor may learn the spurious correlation brought by the backdoor path from X to Z, and thus the backdoor method is used to block the path from X to Z, making it:
    `P(Y|do(Z)) = sigma P(X = x)P(Y|X = x,Z)`
    where `P(X = x)` is known as **Cross-Sampling** and making the whole equation:
    `P(Y|do(X)) = sigma P(Z = z|X) sigma P(X = x)P(Y|Z = z, X = x)`



- Causal Intervention for Weakly-Supervised Semantic Segmentation
  - ▶ Maheep's Notes

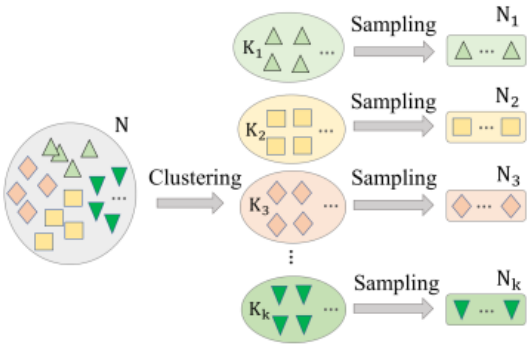
    `C = {c1, c2, c3,.....,cn}` where `n` is the class size to finally compute the equation.

$$M_{t+1} = \sum_{i=1}^{n} \alpha_i c_i P(c_i),\ \alpha_i = softmax\left(\frac{(\mathbf{W}_1 X_m)^T (\mathbf{W}_2 c_i)}{\sqrt{n}}\right),$$



- Confounder Identification-free Causal Visual Feature Learning
  - ▶ Maheep's Notes

    `P(Y|Z = h(x), x̃) = f(Z = h(x))P(x̃)`
    after clustering-then-sampling using the k-mean.



Algorithm 1 Confounder Identification-free Causal Visual Feature Learning

Input: Training dataset $\{(x_i, y_i)\}_{i=1}^{N}$.
Init: learning rate: α, β; model f with parameter θ.
Cluster the training data into K clusters. ▷ Figure 3
while not converge **do**
  Sample M samples from K clusters as a batch.
  Estimate global intervention with $g_\dagger$. ▷ Eq. 8
  Update f with $g_\dagger$ as: $\theta_\dagger = \theta - \alpha g_\dagger$.
  Compute the loss $\mathcal{L}_{CICF}$. ▷ Eq. 11
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{CICF}$.
end while

- Comprehensive Knowledge Distillation with Causal Intervention
  - ▶ Maheep's Notes

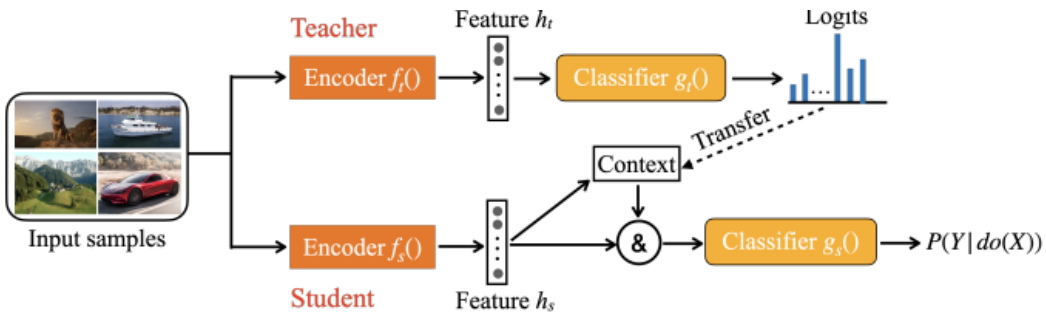
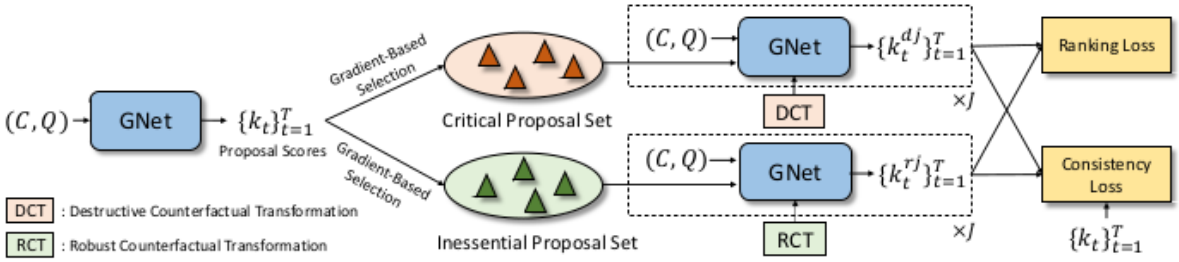
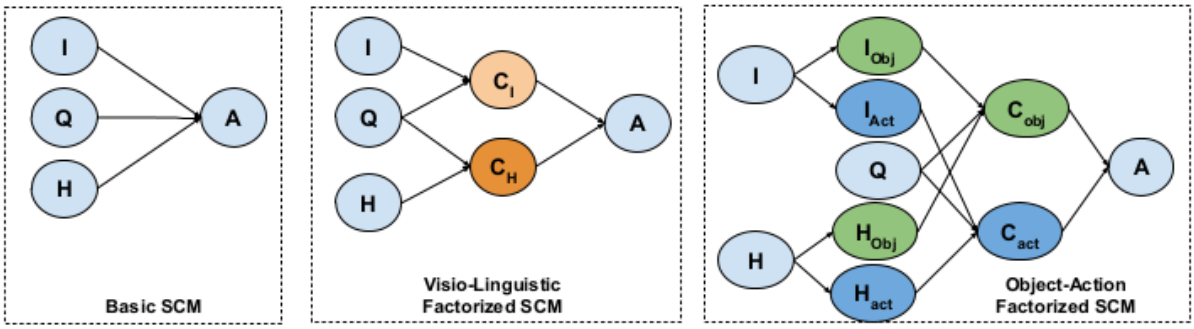


Figure 5: Interventional Distillation. A network can be represented as an encoder $f()$ followed by a linear classifier $g()$ so that teacher $T(X) = g_t(f_t(X))$ and student $S(X) = g_s(f_s(X))$.

- Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding
  - ▶ Maheep's Notes



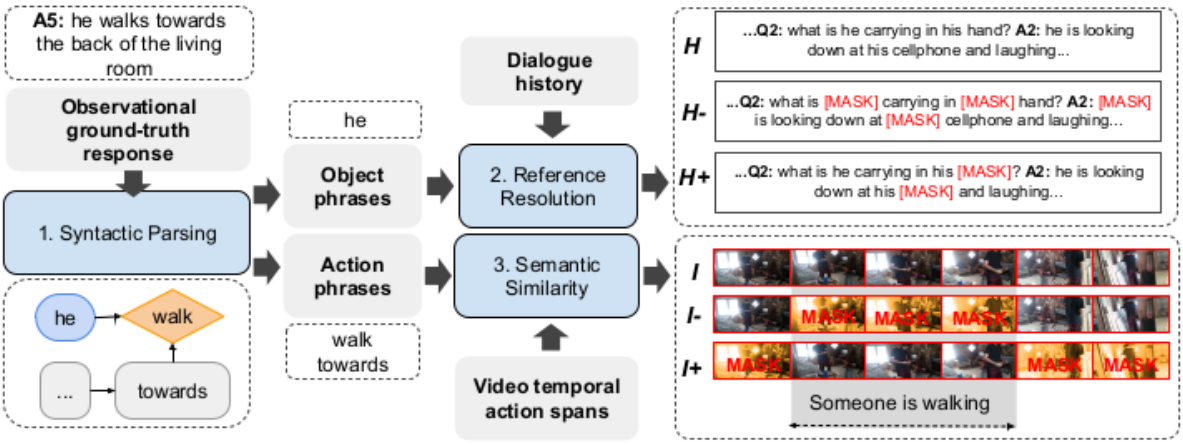
- C_3 : Compositional Counterfactual Constrastive Learning for Video-grounded Dialogues
  - ▶ Maheep's Notes



Also they generate counterfactual scenarios by removing irrelavant objects or actions to create factual data and by removing relevant object or actions, they generate counterfactual data, finally making the equations as:

```
H_t^- = H_{t, obj}^- + H_{t, act}
H_t^+ = H_{t, obj}^+ + H_{t, act}
I^- = I_obj + I_act^-
I^+ = I_obj + I_act^+
```

where `H_t^-` denotes counterfactual dialogue context in instance `t` and `I^-` represents the counterfactual image input.



- COIN: Counterfactual Image Generation for VQA Interpretation
  - ▶ Maheep's Notes

`I'` based on the original image , the latter is concatenated with the attention map `M`, such that the concatenation `[ I; M]` serves as an input to the generator `G`, where the answer is passed

into the `G` so as to create `I'`, where the regions are identified using GRAD-CAM, where the discriminator `D` ensures that image looks realistic and reconstruction loss is used to do miimal changes. The whole process happens as shown in the figure.

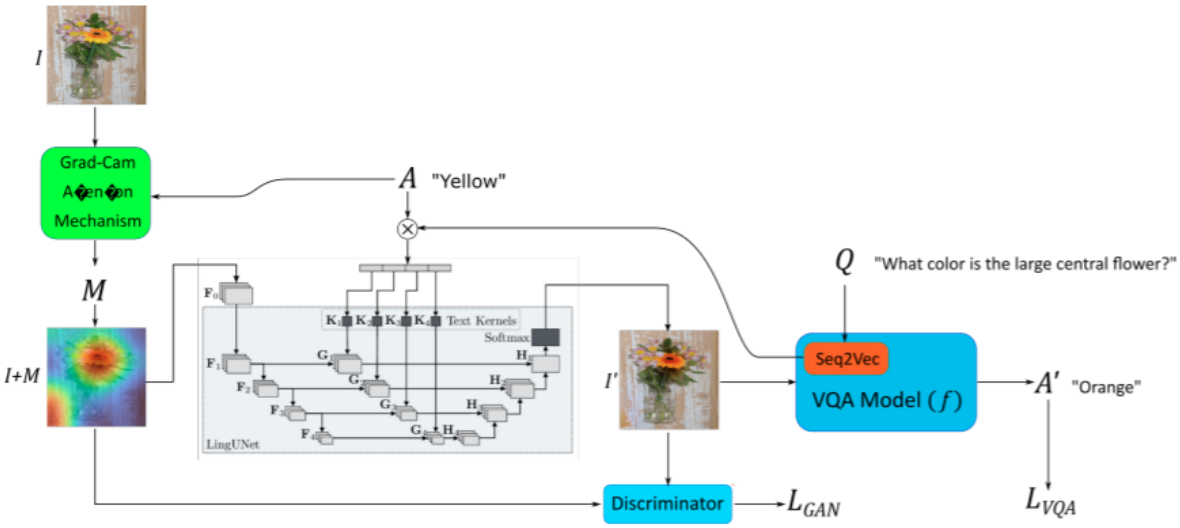
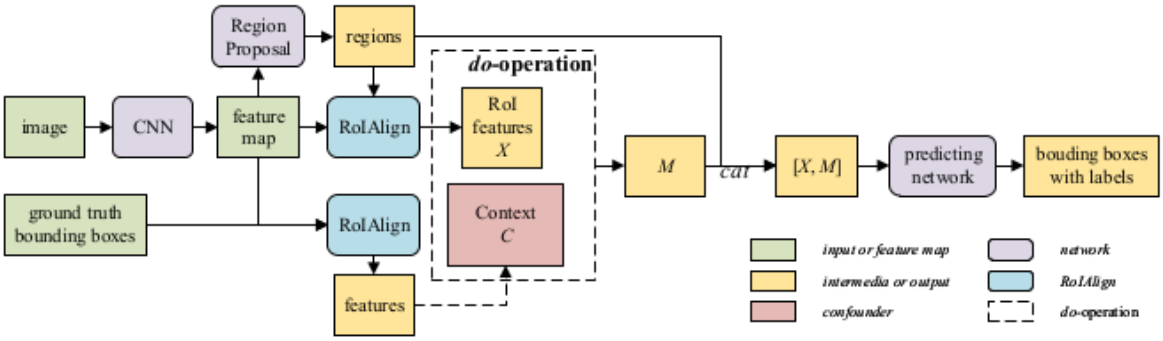


Figure 1. Overview of the proposed architecture inspired by Pan et al. [33]

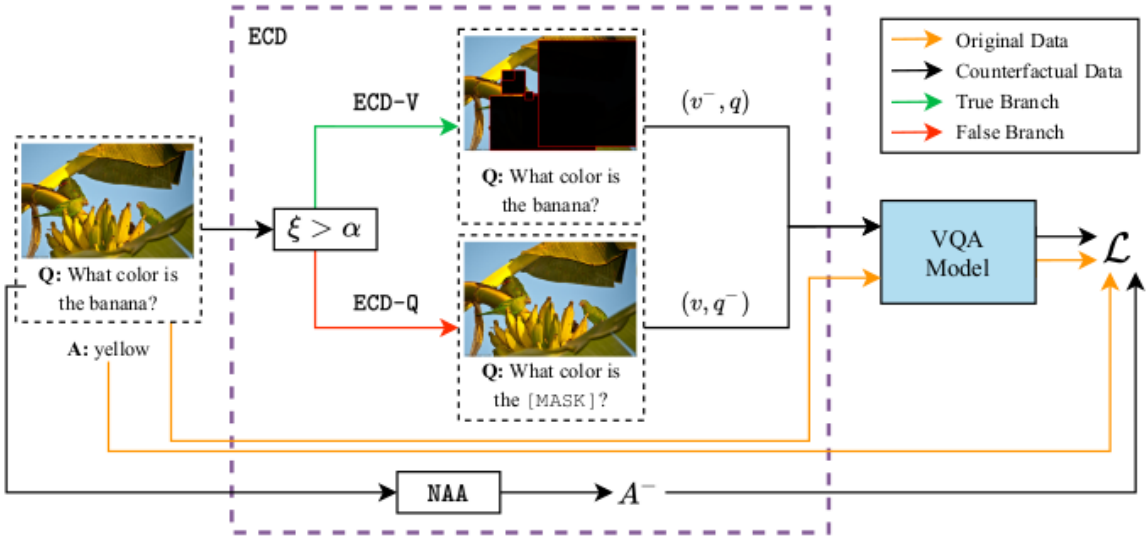
- Causal Intervention for Object Detection
  - ▶ Maheep's Notes

    `P(Y|do(X))` where the author proposes 4 variables namely input `X`, output `Y`, context confounder `C` and mediator `M` affected by both `X` and `C`, where the `C = {c1. c2, ..., cn}` belonging to different `n` categories in the dataset. The output `P(Y|do(X))` is represented as:

    `P(Y|do(X)) = sigma P(c)P(Y|X,M = f(X,c))` where `M` is represented as
    `M = sigma a_i*c_i*P(c_i)`
    where `a_i` is the attention for category specific entry `c_i`.

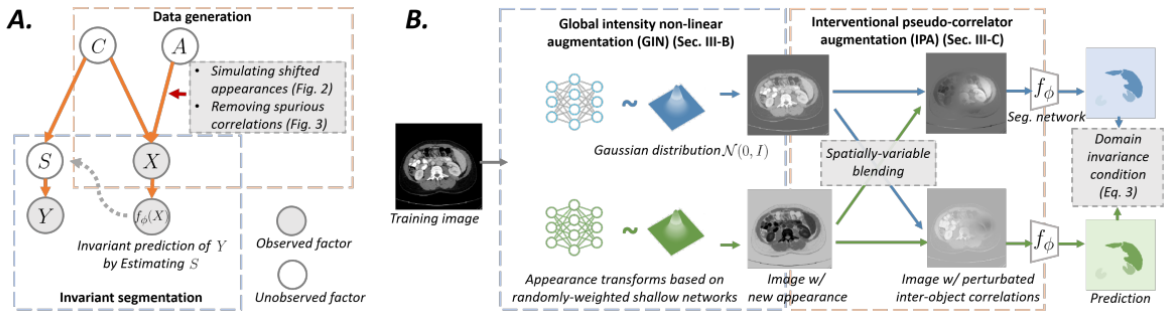


- Efficient Counterfactual Debiasing for Visual Question Answering
  - ▶ Maheep's Notes



- Causality-inspired Single-source Domain Generalization for Medical Image Segmentation
  - ▶ Maheep's Notes

    `do(.)` to remove the confouning nature of on `A` on `S` by transforming the `A` using the `T_i(.)` photometric transformation.
    The psuedo-correlation is proposed so as to deconfound background that is correlted with the output by changing the pixels that correspond to different values are given different values unsupervised fashion. The pseudo-correlation map is impelemnted by using the continous random-valued control points with low spatial frequency, which are multiplied with the `GIN` augmented image.



- Distilling Causal Effect of Data in Class-Incremental Learning
  - ▶ Maheep's Notes

    The paper focuses on to explain the causal effect of these methods. The work proposes to calculate the effect of old data on the current predicition

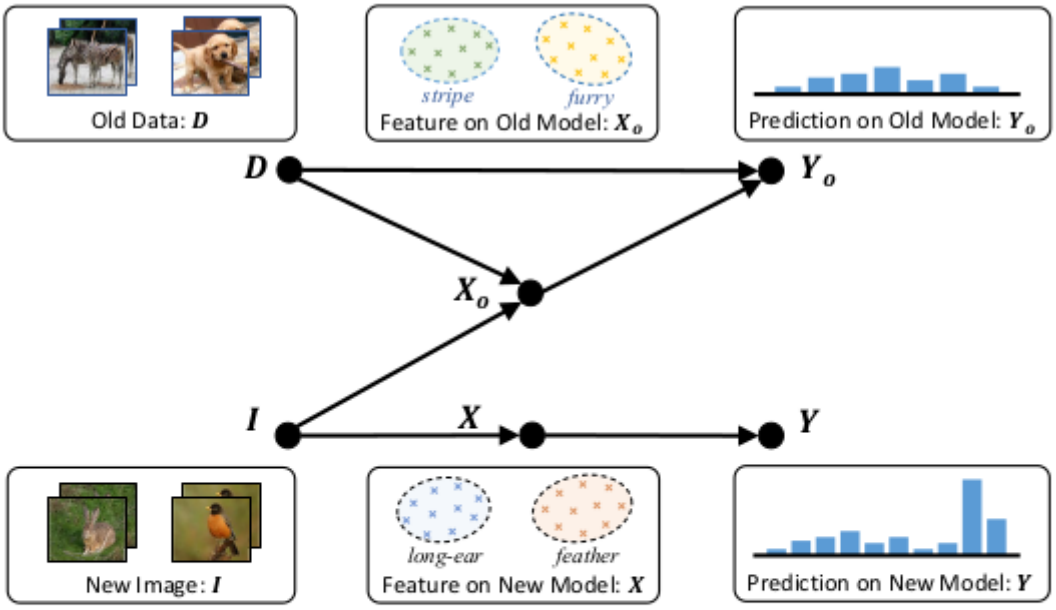
    `Y`, making the equation `Effect = P(Y|D = d) - P(Y|D = 0)`, which comes `0` when the old data has no influence on `Y`, while if we calculate the impact in replay or distillation, will not be `0`. The work proposes to further enhace the replay method by passing down the causal effect of the old data, rather than the data. Therefore making the whole process computationally inexpensive by conditioning on `Xo`, i.e. the old data representaiton and therefore making the equation:

`Effect = P(Y|I, Xo)(P(I|Xo, D = d) - P(I|Xo, D = 0))`

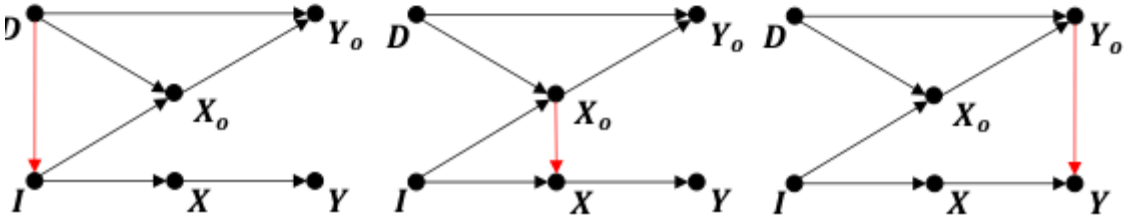
further defining it as:

`Effect = P(Y|I, Xo)W(I = i, Xo, D)`

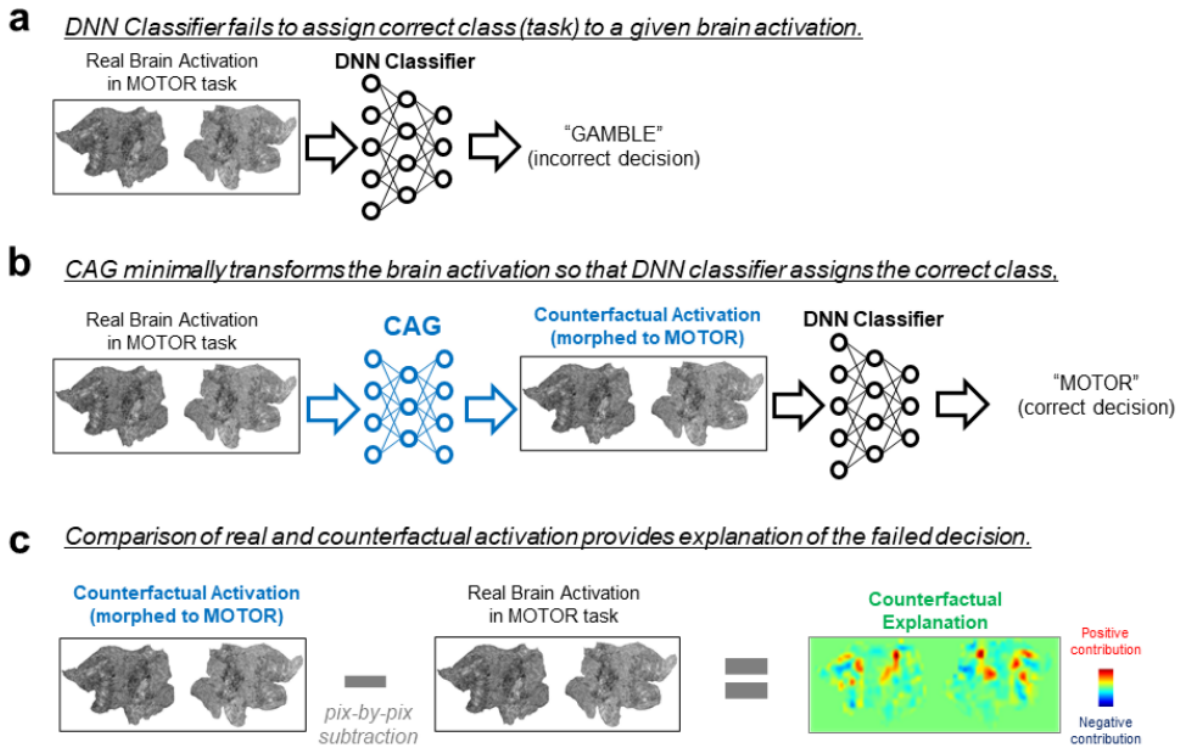
The paper aims to increase the value of `W(.)` expression as it depends the close similarity between the representation of the similar image in old model and new model.



(a) The Forgetting of Class-Incremental Learning



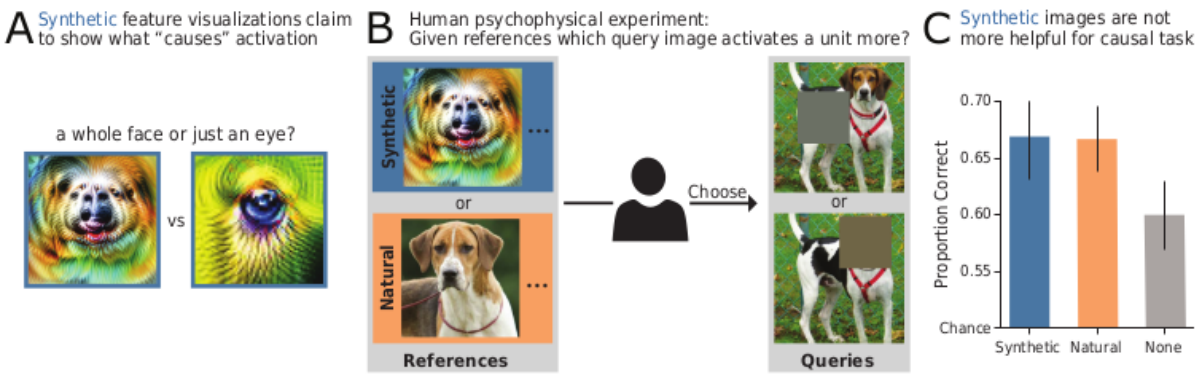
- Counterfactual Explanation of Brain Activity Classifiers using Image-to-Image Transfer by Generative Adversarial Network
  - ▶ Maheep's Notes



- How Well do Feature Visualizations Support Causal Understanding of CNN Activations?
  - ▶ Maheep's Notes

    1.) **Synthetic Reference** : These are image that are generated from optimized result of feature visualization method.
    2.) **Natural Reference** : Most strong actiavted samples are taken from the dataset.
    3.) **Mixed Reference** : 4 synthetic and 5 Natural refernce are taken, to take the best of both worlds
    4.) **Blurred Reference** : Everything is blurred, except a patch.
    5.) **No Reference** : Only query image is given and no other image.

    The author concludes the research by concluding that the performance of humans with visualization and no visualization did not have very significant differences.



- CausalAF__Causal_Autoregressive_Flow_for_Goal_Directed_Safety_Critical
  - ▶ Maheep's Notes

    Behavioural Graph unearths the causality from the **Causal Graph** so as to include in the generated samples. This is done using two methods namely:

1.) **COM** : It maintians the **Q** , to ensure that the cause is generated in terms of nodes only after the effect. It is also noted that the node have many parents, therefore the node is considered valid only when all of it's parents have been generated.
2,) **CVM** : The correct order of causal order is not sufficient for causality therefore CVM is proposed so as to only consider the nodes when the information of it's parents are available and the information only flow to a node from it's parents.

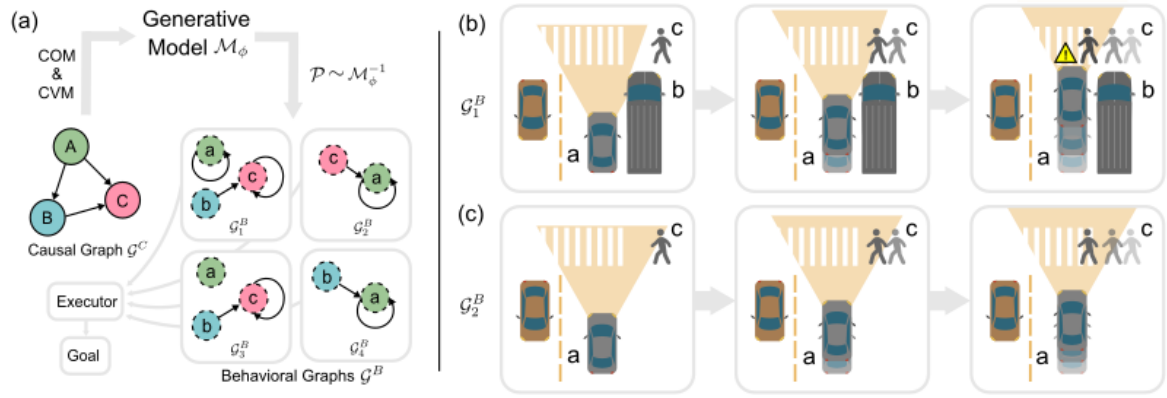


Figure 1: **Left.** Diagram of goal-directed generation and *CausalAF*. **Right.** Two examples obtained by executing two Behavioral Graphs to show the causation behind scenes. **(b)** is safety-critical because the vision of autonomous vehicle a is blocked by vehicle b. In contrast, **(c)** is safe for vehicle a since there is no vehicle b blocking the vision of a. In general, b is the cause of the collision.

- Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-identification
  - ▶ Maheep's Notes

    `A = {A1, A2, A3,....., An}`, the attention maps are used to extract the respective feature from the image. `hi = gamma(X*Ai)`, where all the `hi` are normalized to get the `h = normalize(h, h2,...., hn)` which is used to predict. 2.) The attention is intervened to get the effect on the output of the model, i.e.
    `Y_effect = E[Y(A = A, X = X) - Y(do(A = bar(A))), X = X]`
    It is expected to achieve two-conditions using this method:
    a.) The attention model should improve the prediction based on wrong attentions as much as possible, which encourages the attention to dis- cover the most discriminative regions and avoid sub-optimal results
    b.) The prediction based on wrong attentions is penalized, which forces the classifier to make decision based more on the main clues instead of the biased clues and reduces the influence of biased training set.

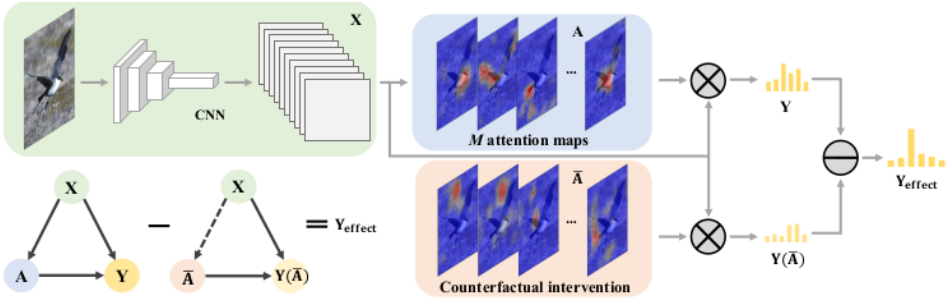
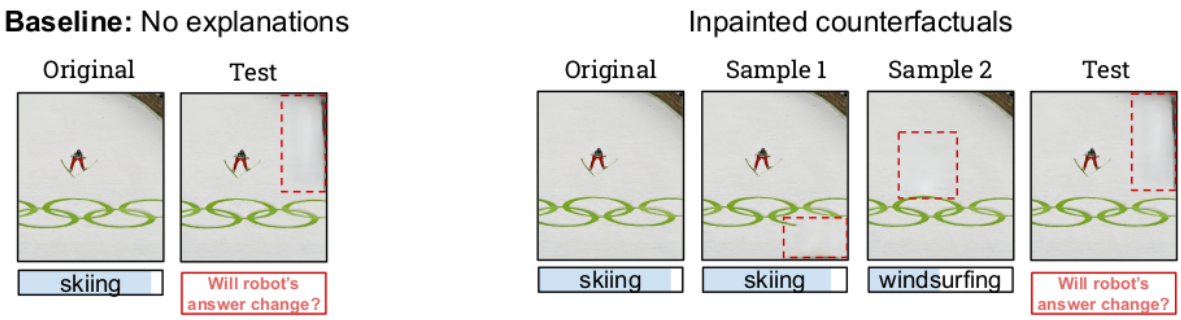


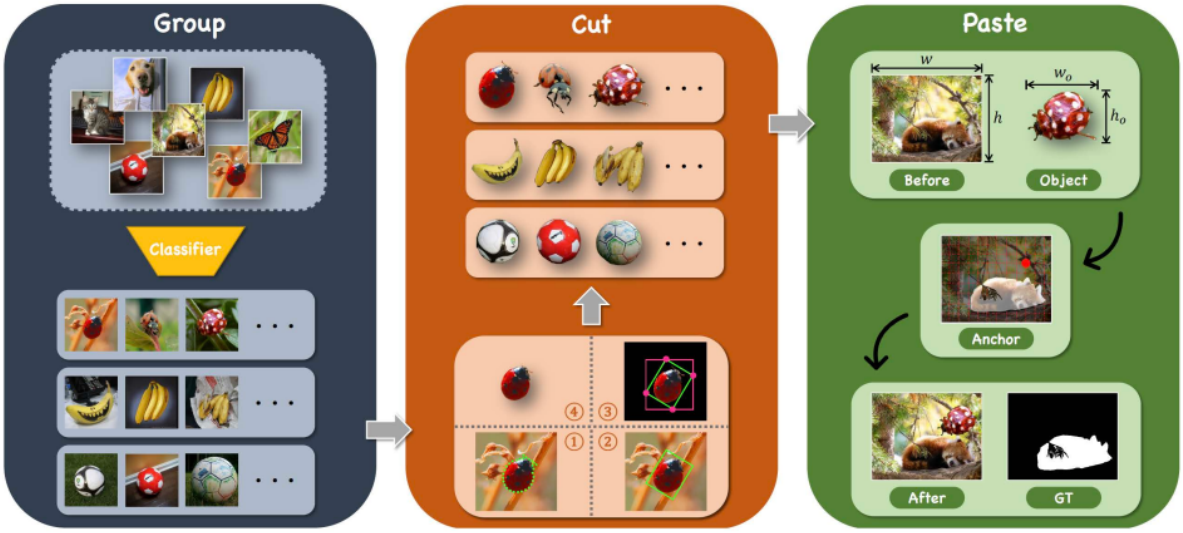
Figure 3: The overall framework of our CAL method. We first apply the counterfactual intervention for original attention by replacing with random attentions. Then, we subtract the counterfactual classification results from original classification to analyze the effects of learned visual attention and maximize them in the training process.

- Improving Users' Mental Model with Attention-directed Counterfactual Edits
  - ▶ Maheep's Notes



- Free Lunch for Co-Saliency Detection: Context Adjustment
  - ▶ Maheep's Notes

    group-cut-paste method to imporve the data distribution. GCP turns image `I` into a canvas to be completed and paint the remaining part through the following steps:
    (1) classifying candidate images into a semantic group Z (e.g., banana) by reliable pretrained models
    (2) cutting out candidate objects (e.g., baseball, butterfly, etc.)
    (3) pasting candidate objects into image samples as shown in the figure below.
    To make the process more robust the author proposes to have three metrics, namely:
    a.) **Abduction**: In the new generated data the co-saliency image should remina unchanged.
    b.) **Action**: The mask sould remain unchanged from the GT of the image and should be optimal for it's value.
    c.) **Prediction**: The probability distribution of the image should remian unchanged.



- Counterfactual Explanation Based on Gradual Construction for Deep Networks
  - ▶ Maheep's Notes

    1.) **Masking Step**: It mask the appropriate region of the image, to which the model pays most attention, extracted using the gradients.
    2.) **Composition Steps**: It perturbs the regions minimally so as to change the logits to the target class.

Algorithm 1: Gradual construction

Input:
- $X \in \mathbb{R}^d$: an input data
- c_t: a target class
- τ: the desired classification probability for the target class
- σ: the number of iteration

Initialization:
- Mask $M \in \mathbb{R}^d$ and $M_i = 0 \ \forall i$
- Composite $C \in \mathbb{R}^d$ and $C_j \sim N(0,1) \ \forall j$
- Perturbed data $X' = (1 - M) \circ X + M \circ C$
- The number of perturbed features $n = 1$

While $f_{c_t}(X') < \tau$:

1) Masking step

$i^* \leftarrow$ an index of the n highest value in $|\bigtriangledown f_{c_t}(X)|$

$M_{i^*} \leftarrow 1$

2) Composition step

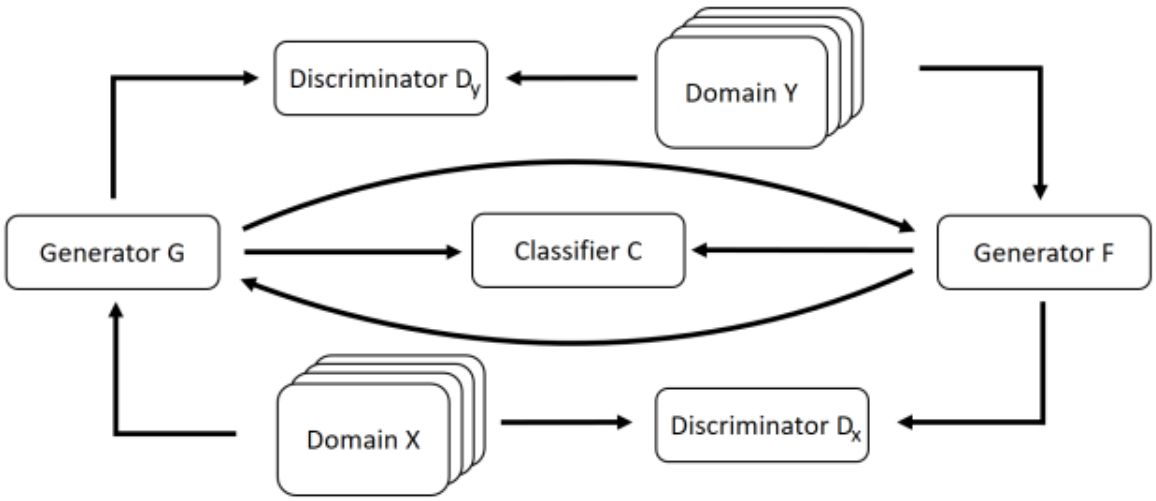
for $m = 1$ **to** σ **do**

$$C \leftarrow \arg\min_C \left\| \sum_{k=1}^{K} \left(f'_k(X') - \frac{1}{N} \sum_{i=1}^{N} f'_k(X_{i,c_t}) \right) \right\|_2 + \lambda \left\| X' - X \right\|_2$$

$n \leftarrow n + 1$

Output: X'

- GANterfactual - Counterfactual Explanations for Medical Non-Experts using Generative Adversarial Learning
  - ▶ Maheep's Notes



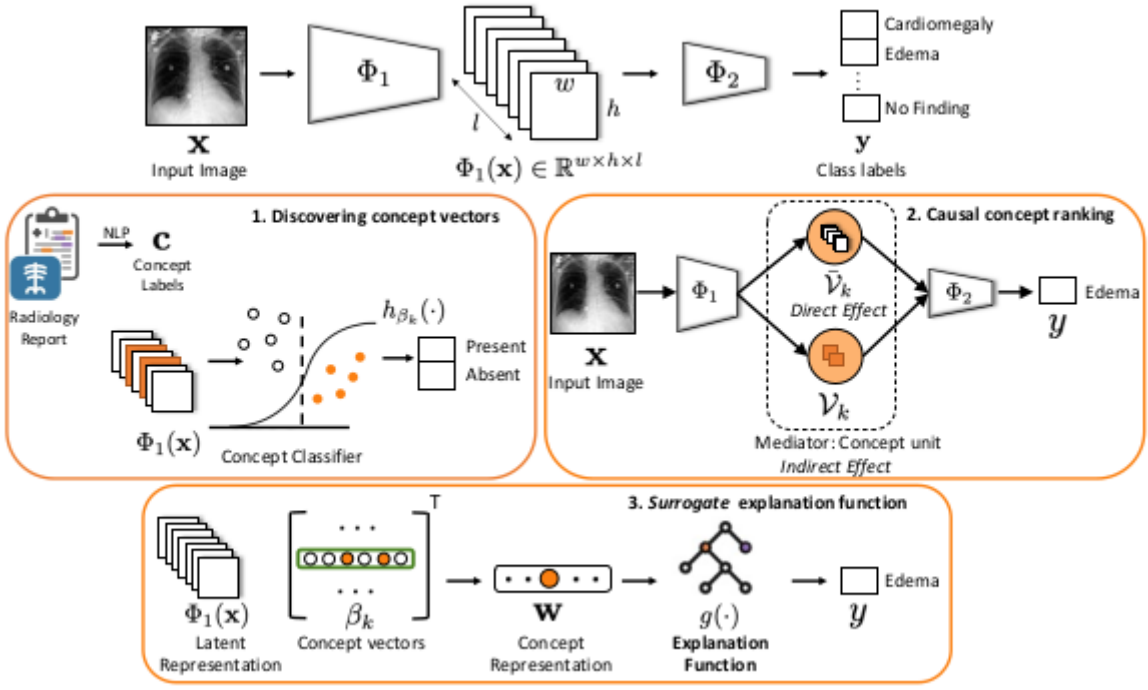
- Using Causal Analysis for Conceptual Deep Learning Explanation

  - ▶ Maheep's Notes

    `phi1(.)` and `phi2(.)`, where the `phi1(.)` gives different concept in terms of features and `phi2(.)` do prediction. The output of `phi1(.)` gives a vector of `lbh` dimension with each unit having a binary prediction, i.e. if concept is present or absent.
    2.) **Causal concept ranking**: A counterfactual `x'` for the input `x` is generated for causal inference using a cGAN, where the concepts are denoted with `Vk(x)` and the left over hidden units are denoted by `bar(Vk(x))` and the effect is measured by:

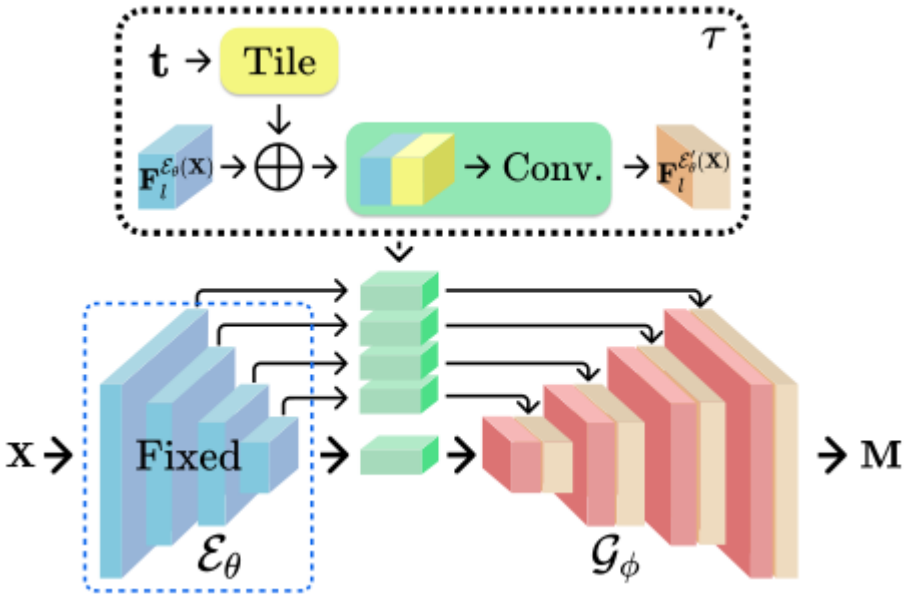
```
A = phi2(phi1(do(Vk(x)), bar(Vk(x'))))
B = phi2(phi1(Vk(x), bar(Vk(x))))
Effect = E[A/B - 1]
```

    3.) **Surrogate explanation function**: A function `g(·)` is introduced as a decision tree because many clinical decision-making procedures follow a rule-based pattern, based on the intial classifier `f(.)` based on the logits produced for different concepts.

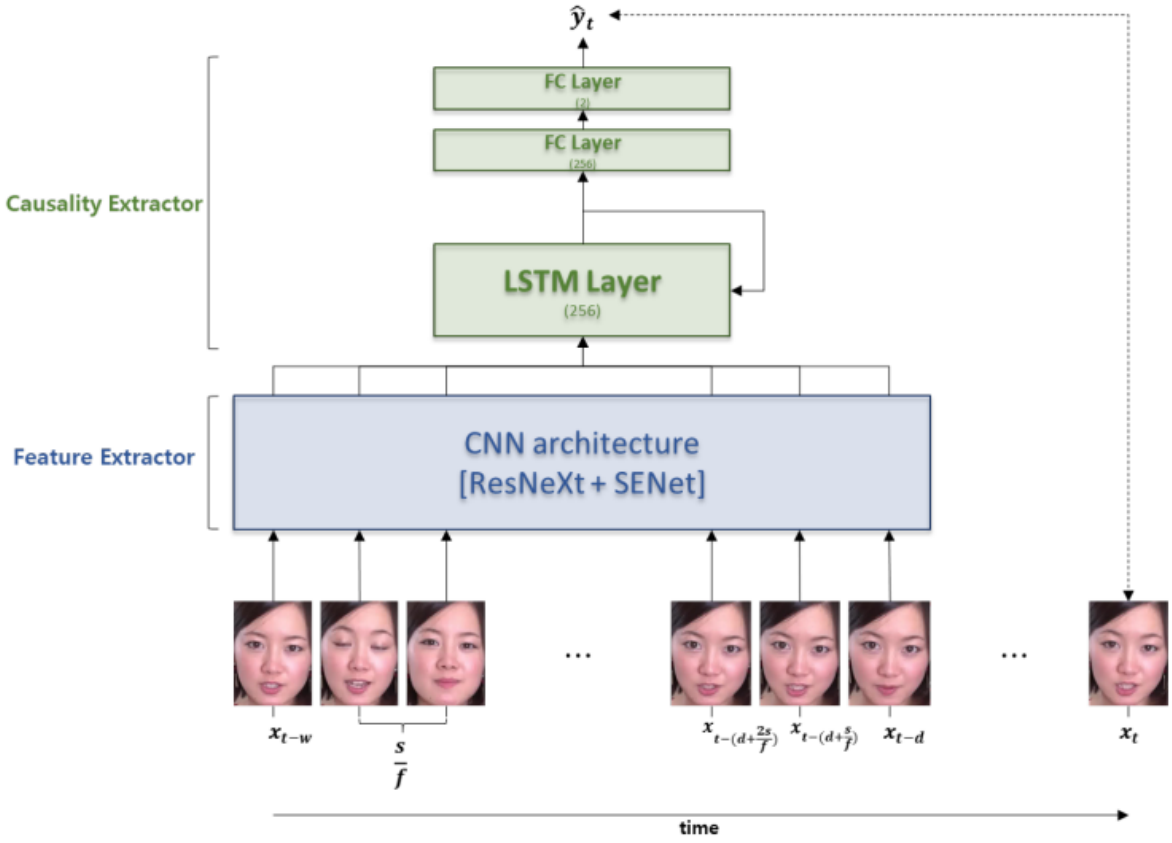


- Learn-Explain-Reinforce: Counterfactual Reasoning and Its Guidance to Reinforce an Alzheimer's Disease Diagnosis Model

  - ▶ Maheep's Notes



- Causal affect prediction model using a facial image sequence
  - ▶ Maheep's Notes



- iReason: Multimodal Commonsense Reasoning using Videos and Natural Language with Interpretability
  - ▶ Maheep's Notes

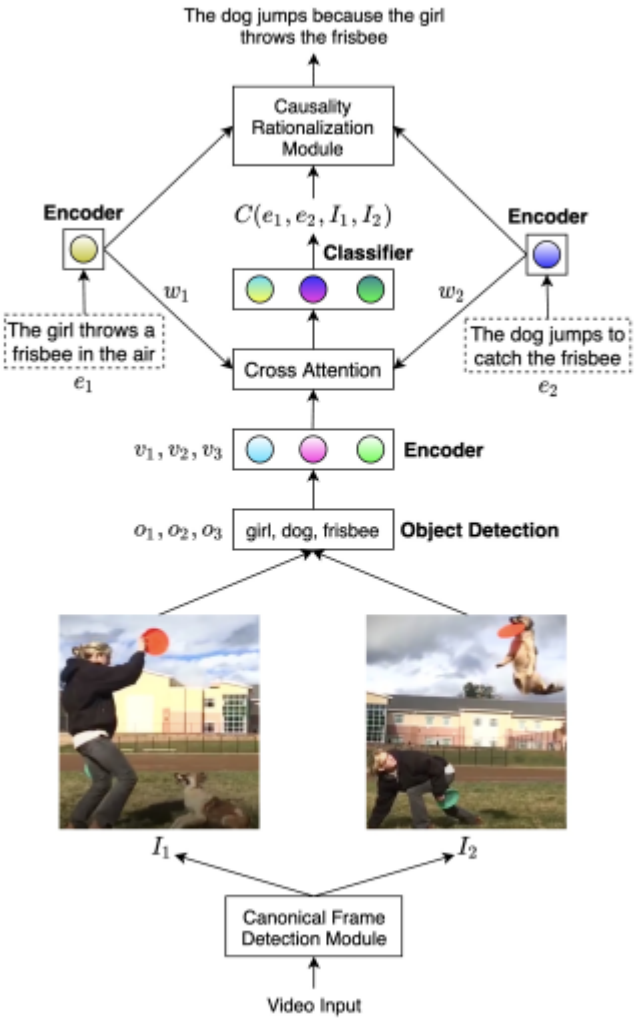


Figure 1: Architectural overview of iReason.

The CFDM module localize the said events in the video and outputs a pair of images, i.e.

`I1` and `I2` from the video. The aim is to infer causality from the event in `I1` into `I2`. The **Causality Rationalization Module** outputs a string explaining the commonsense reasoning using natural language for causal events `e1` from `I1` and `e2` from `I2`.

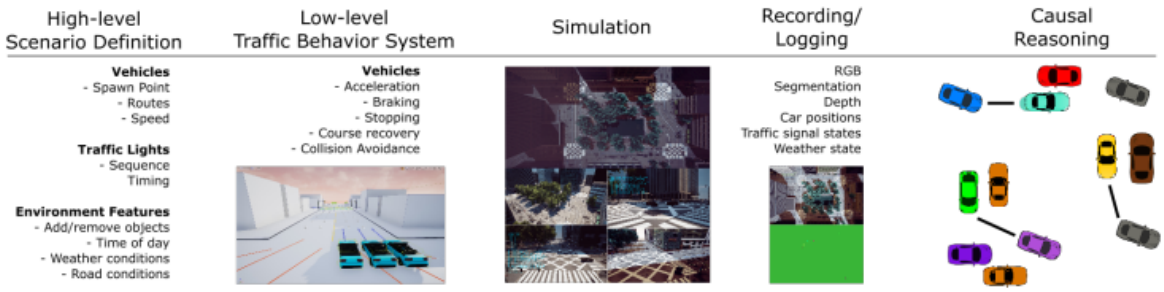
- CausalCity: Complex Simulations with Agency for Causal Discovery and Reasoning
  - ▶ Maheep's Notes

    The author introduces two main things that totallly build up the whole simulated environment, i.e.

    JSON file that contains the vehicles that should be present, their start location, and high-level features which are flexible and can be regulated using the **Python API**. To make the whole enviroment 6 types of varibales are introduced, namely:
    1.) **Environment Features**: It contains the information about the basic objects in the environment like trees, pole, etc.
    2.) **Vehicles**: It contains the vehicle positions, their velocities and information about the collision.
    3.) **Traffic Lights**: It contains the information where the traffic lights will be and how will they react at different time frame.

4.) **Environment**: It contains the information about the weather, from which the confounders can be easily added.
5.) **Views/Cameras**: It has the ability where to place the camera for recording, therefore providing the dataset with third person or bird eye view.
6.) **Logging**: The log of different vehicles and state of traffic lights are recorded in it. Although other things can also be included.

Using the author prepares two types of dataset:
a.) **Toy Dataset**: It contains no confounders, agency but only causal relationship.
b.) **CausalityCity**: It contains confounders, agency and also causal relationship.



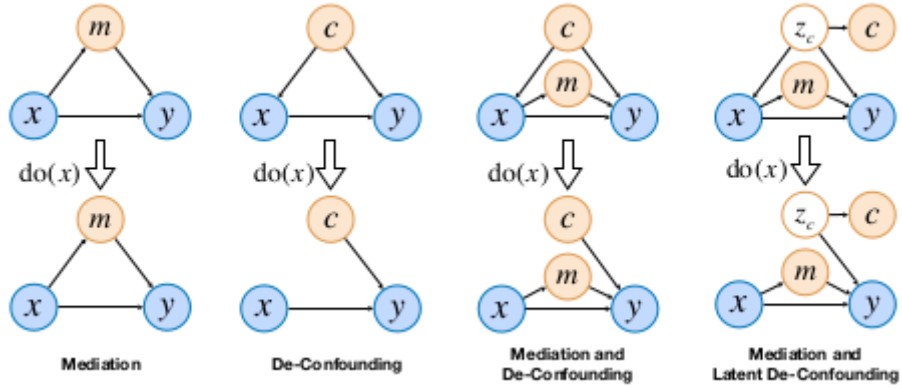
- Driver-centric Risk Object Identification
  - ▶ Maheep's Notes

    Perception, which represents different embeddings of the objects present, **Comprehension** which evaluates the interaction between the driver and thing or stuff using the Ego-Thing Graph and Ego-Stuff Graph, where Ego-Thing Graph have the embedding of how the driver react with the things such as the car, person, bicycle and the Ego-Stuff Graph have the embedding of how the driver reacts with the Stuff in the envionment such as roads, footpath, and Traffic Sign. The last module is of **Projection** which is used to predict the future forecasts.

    The Causal reasoning module is added to the model so as to augment the data only in "GO" scenarion, i.e. no risk objects are present to remove the non-causal features by randomly selecting top k ransdom objects. It is also used in "STOP" scenarios, to identify the risk object identification by using the same intervention maethod of inpainting. The "GO" score is computed by removing the different object and the one object with removal that gives the highest "GO" score is identified as the risk object.

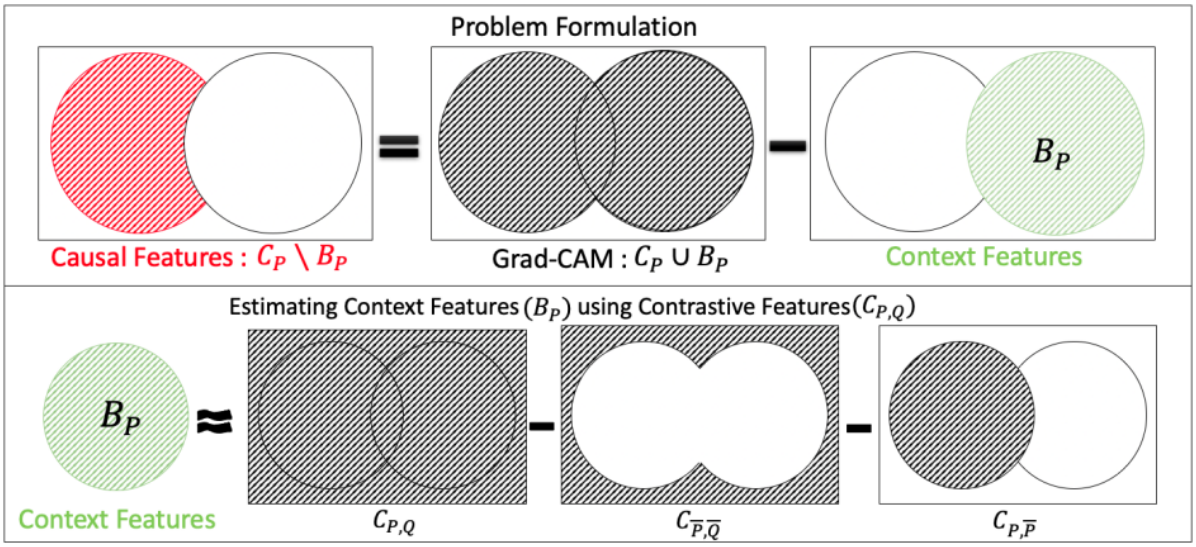
- Dependent Multi-Task Learning with Causal Intervention for Image Captioning
  - ▶ Maheep's Notes

    `m` and the proxy confounder `c` to eradicate the real confounder `z_c`. In these type of systems it is to be considered that the mediator is not affected by the counfounder after the intervention.



- Structure by Architecture: Disentangled Representations without Regularization
  - ▶ Maheep's Notes
- EXTRACTING CAUSAL VISUAL FEATURES FOR LIMITED LABEL CLASSIFICATION
  - ▶ Maheep's Notes

    C, separating them from context features B while computed from Grad-CAM using the Huffman encoding which increases the performance by 3% in terms of accuracy and also retains 15% less size in bit size.
    The author implements it by arguing that given the just features G = C U B are given. By taking the analogy of the sets given below, the author extracts B as given in the following equations below:



    C_p = G_p - B_p,.....(1) i.e. for prediciton p
    B_p = C_(p,q) - C_(bar(p),bar(q)) - C_(bar(p),p) .....(2)
    which denotes the following things:
    C_(p,q): "Why p or q?"
    C_(bar(p),bar(q)): "Why neither P nor Q"
    C_(bar(p),p): "Why not P with 100% confidence?"

    Therefore (1) can be eaily be obtained after substituting the value of (2) in it.

- ALIGN-DEFORM-SUBTRACT: AN INTERVENTIONAL FRAMEWORK FOR EXPLAINING OBJECT DIFFERENCES

  - ▶ Maheep's Notes

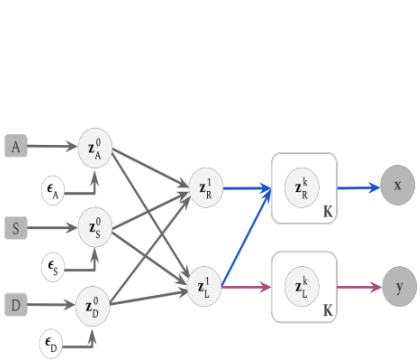
    X_s converting into the target object image X_t, by modifying it and quantifying the parameters via changing it's **affnity** by changing the scaling s, translation Δt and in-plane rotation Δθ. **Shape** acts as the second parameter by which the image is changed. The transformation takes place in the same order as if shape is changed before that it will also have the effect of changing the pose of the image. **Subtract** act as the third module to change the image via removing the background using a segmentaion model to see the apperance difference using MSE.

- Translational Lung Imaging Analysis Through Disentangled Representations

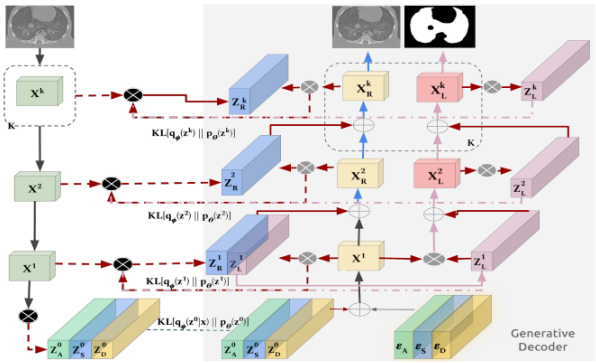
  - ▶ Maheep's Notes

    The author implements it by considering 3 factors for generating and masking the image, namely: animal model,

    A, the realtive position of axial slice, S and estimated lung damage, Mtb, via the hierarchy at different resolution scales k. By using the Noveau VAE to extract the latent space z variables to generate the mask y and image x.



(a) Direct Acyclic Graph (DAG)



(b) Summarized Architecture

- CRAFT: A Benchmark for Causal Reasoning About Forces and inTeractions

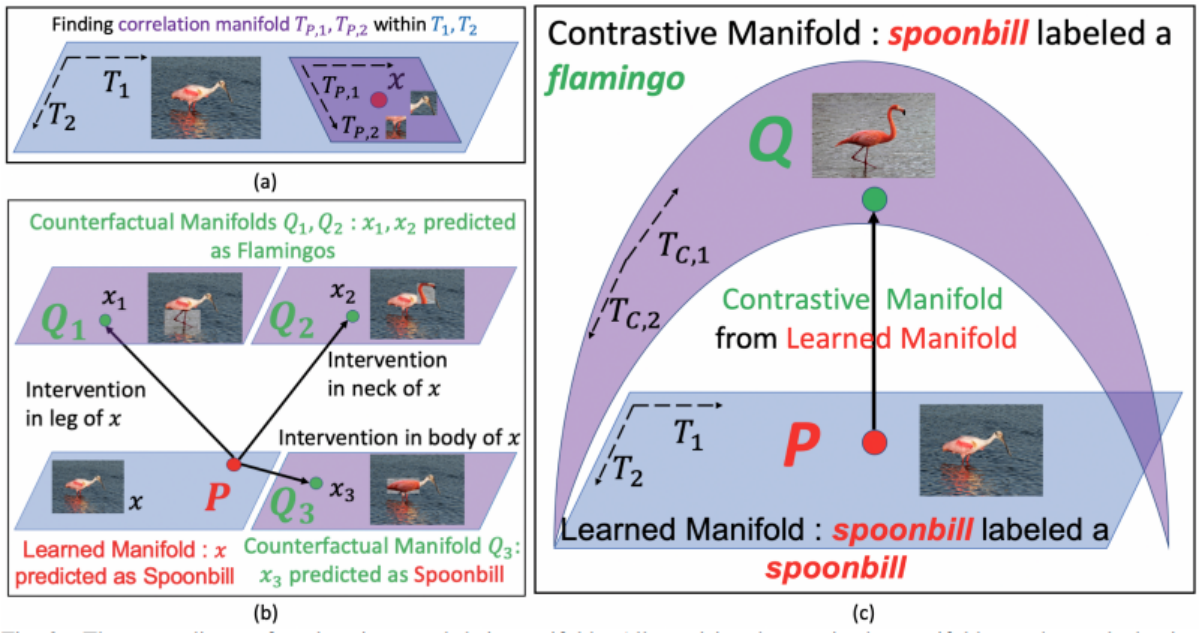
  - ▶ Maheep's Notes

    1.) **Descriptive Questions** : It requires extracting the attributes of objects, especially those involving counting, need temporal analysis as well
    2.) **Counterfactual Questions** : It requires understanding what would happen if one of the objects was removed from the scene. For ex: *"Will the small gray circle enter the basket if any of the other objects are removed?"*

3.) **Causal Questions** : It involves understanding the causal interactions between objects whether the object is causing, enabling, or preventing it.

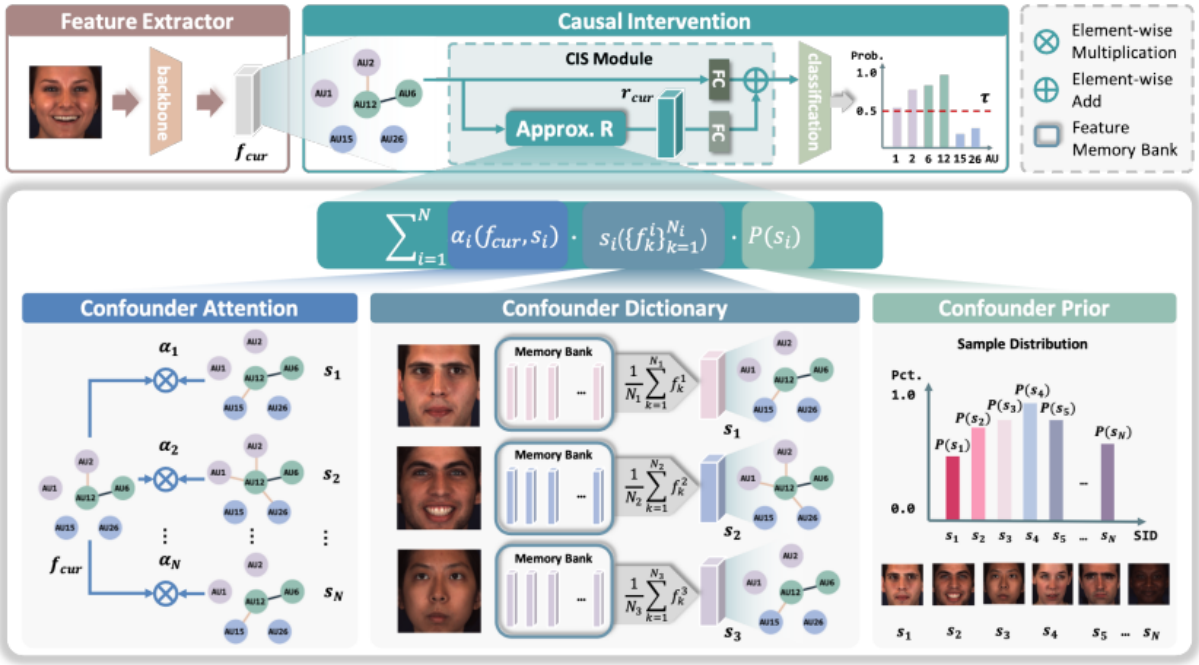
- Explanatory Paradigms in Neural Networks
  - ▶ Maheep's Notes

    Abductive Reasoning from the three reasoning methods, including **Deductive Reasoning** and **Inductive Reasoning**. The author explains the **Abductive Reasoning** hypothesises to support a prediction and if seen abstractly defines it into three major fields, explained clearly by taking manifold into the picture, dealing with:
    1.) **Observed Correlation**: It majorly deals with the question *Why P?*, where *P* is a class. The goal here is find all the dimensions of the manifold, denoted by `T_f` from the constructerd manifold, denoted by `T` that justifies the class *P* classification from the network, denoted by `M_cu(.)`
    2.) **Observed Counterfactual**: It majorly deals with the counterfactual question, i.e. *What if not P?*. It deals with interventions so as to change the direction of some of the dimensions by intervention using `do(.)` calculus to indetify the most non-trivial features a specific attribute of *P* denoted by `M_cf(.)`
    3.) **Observed Contrastive Explanations**: It majorly deals with the counterfactual question, i.e. *What P rathre than Q?*. It deals with interventions so as to change the direction of some of the dimensions to chenge the prediciton of network from `P` to `Q`, to identify the most non-trivial features separating class from *P* and *Q* denoted by `M_ct(.)`
    The authors discusses **Probabalistic Components of Explanations** that can take into consideration the questions defined above and make explanation more substatial by:
    `M_c(x) = M_cu(x) + M_ct(x) + M_cf(x)`
    Besides this the author discusses about the **Contrast-CAM**, **Counterfactual-CAM**, **Grad-CAM** technique which is generally used for observing Observed Correlations. The **Counterfactual-CAM** is used for **Observed Counterfactual** negates the gradient to decrease the effect of the predicted class resulting in highlighted regions in case when the object used to make the decision were not present intially. The **Contrast-CAM** is used for third scenario of **Observed Contrastive Explanations** where a loss between class *P* and *Q* is constructed to backpropogate it and find contrastive features.



- A Closer Look at Debiased Temporal Sentence Grounding in Videos: Dataset, Metric, and Approach
  - ▶ Maheep's Notes
- Causal Intervention for Subject-deconfounded Facial Action Unit Recognition
  - ▶ Maheep's Notes

    X, Subject(Confounder) S, Latent Representation R and output Y, where the author eradicates the effect of S on X. The author implements it by having three modules:
    1.) **Attention Module**: It takes the attention of the extracted feature and the different AU for each Subject which are computed by taking the average of all the samples of the Subject, denoted by `s_i`
    2.) **Memory Module**: It consist `s_i` as defined above
    3.) **Confounder Priors**: It consist of the sample distribution of different `s_i` by taking the number of (samples in that subject)/(total samples)



- Causal Scene BERT: Improving object detection by searching for challenging groups of data
  - ▶ Maheep's Notes

    weather patterns, *Vehicle types* and *Vehicle positioning*. The author harnesses the Simulation and **MLM**(Masked Language Model) to apply causal intervention so as to generate counterfactual scenarios while **MLM**, acts as a Denoising Autoencoder to generate data near true distribution. The different tokens represent different groups such as *weather*, *agent asset*, *rotations*, etc. and are masked to generate counterfactual image. The author uses the score function `f(phi, I, L)` where `phi` is the model, `I` is the image and `L` is the label. The score function is used to identify the vulnerable groups using the `rho` function:
    `rho` = `f(phi, I', L')` - `f(phi, I, L)`
    if `|rho| >= t`, where `t` is the threshold, which defines if the `rho` is very negative or positive then the modified group is vulnerable.