A Rigorous Study of Integrated Gradients Method with Extensions to Internal Neuron Attributions

International Conference on Machine Learning, 2022

Daniel Lundstrom, Tianjian Huang, & Meisam Razaviyayn

Contributions

Contributions

- Identify problems with uniqueness claims in [Sundararajan 2017], [Xu et al, 2020], [Sundararajan & Naomi, 2020]

Contributions

- Identify problems with uniqueness claims in [Sundararajan 2017], [Xu et al, 2020], [Sundararajan & Naomi, 2020]
    - Rigorously establish the claims with an additional axiom: nondecreasing positivity.

Contributions

- Identify problems with uniqueness claims in [Sundararajan 2017], [Xu et al, 2020], [Sundararajan & Naomi, 2020]
    - Rigorously establish the claims with an additional axiom: nondecreasing positivity.
- Study when IG is or may fail to be Lipschitz continuous.

Contributions

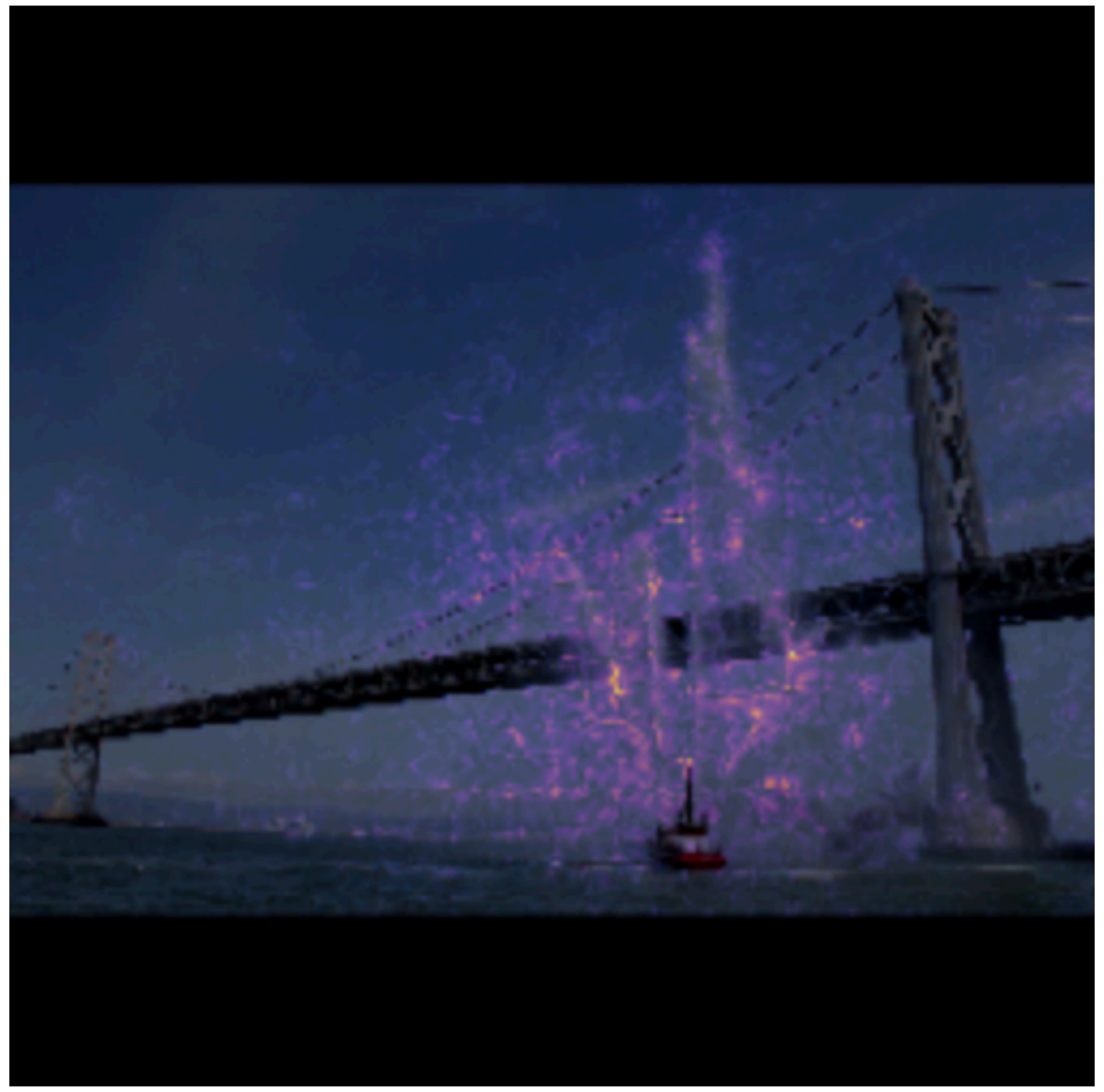
- Identify problems with uniqueness claims in [Sundararajan 2017], [Xu et al, 2020], [Sundararajan & Naomi, 2020]
  - Rigorously establish the claims with an additional axiom: nondecreasing positivity.
- Study when IG is or may fail to be Lipschitz continuous.
- Introduce axioms when IG has a distribution of baselines.

Contributions

- Identify problems with uniqueness claims in [Sundararajan 2017], [Xu et al, 2020], [Sundararajan & Naomi, 2020]
    - Rigorously establish the claims with an additional axiom: nondecreasing positivity.
- Study when IG is or may fail to be Lipschitz continuous.
- Introduce axioms when IG has a distribution of baselines.
- Introduce region-targeting attribution method for internal neurons.

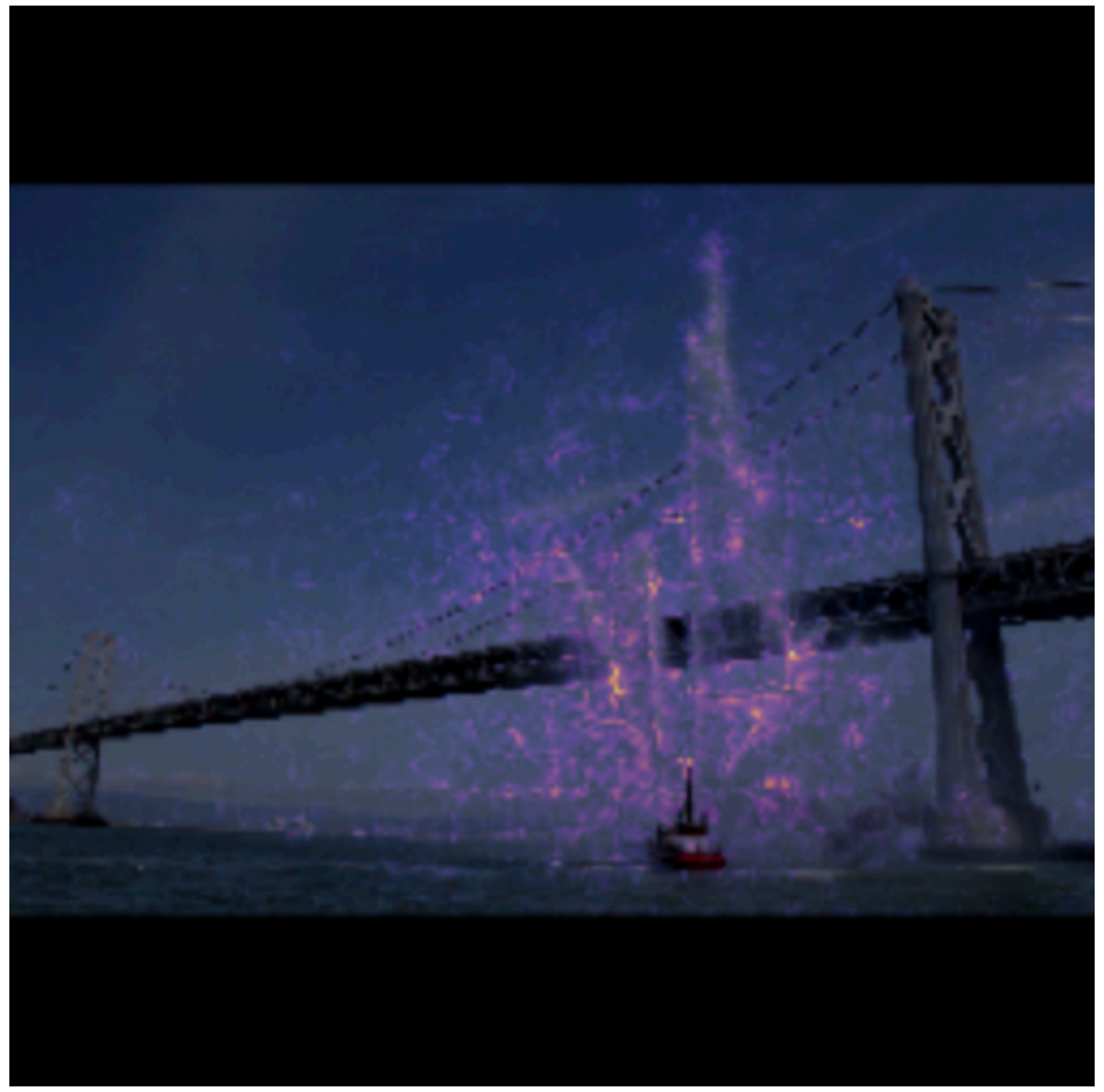
The Method of Integrated Gradients(IG)

The Method of Integrated Gradients(IG)

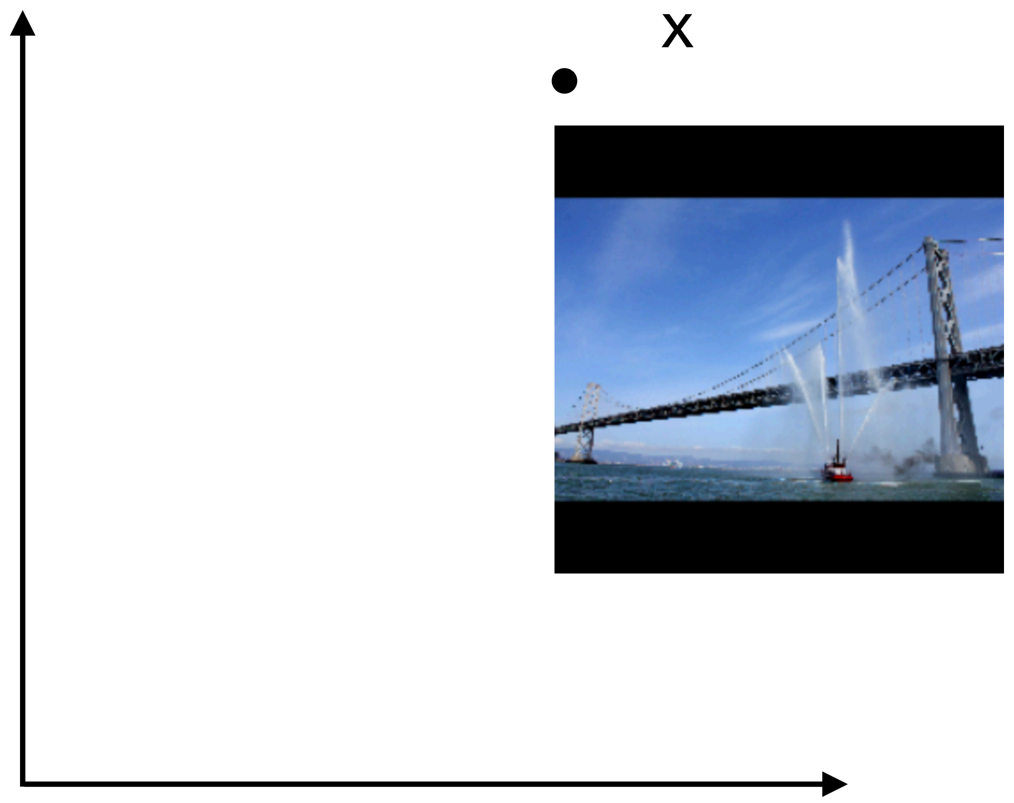


Fireboat & IG

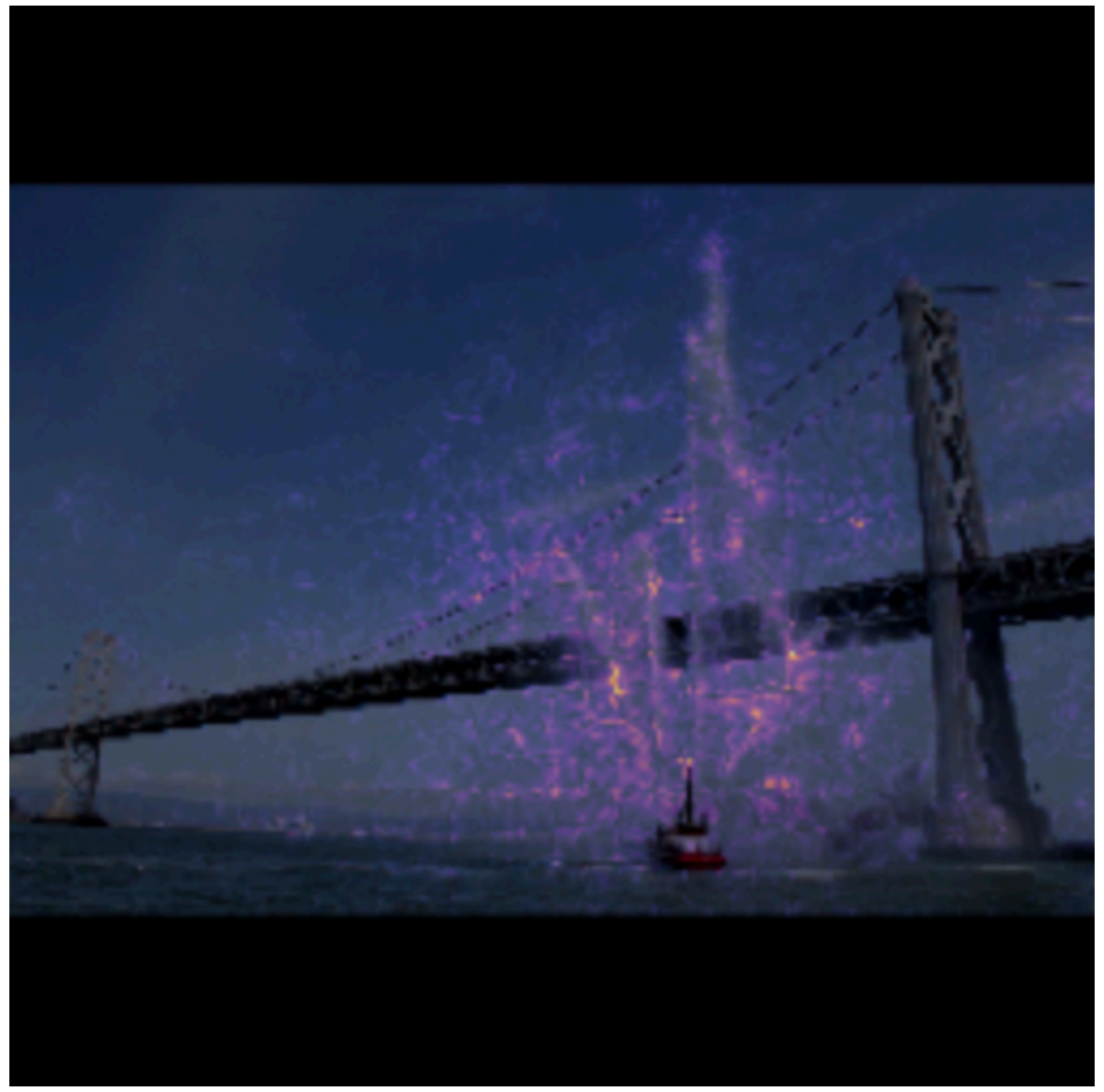
The Method of Integrated Gradients(IG)



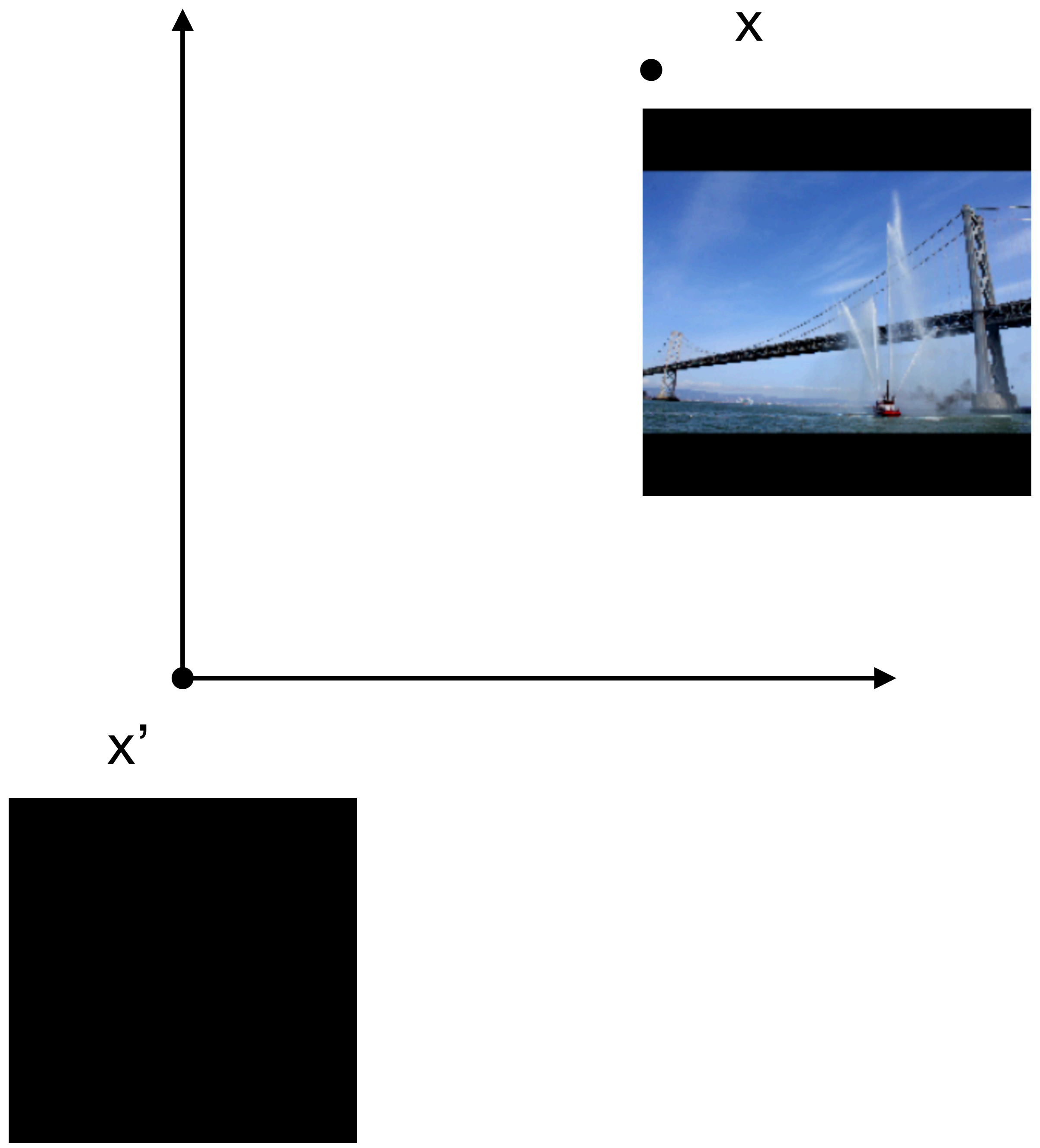
Fireboat & IG



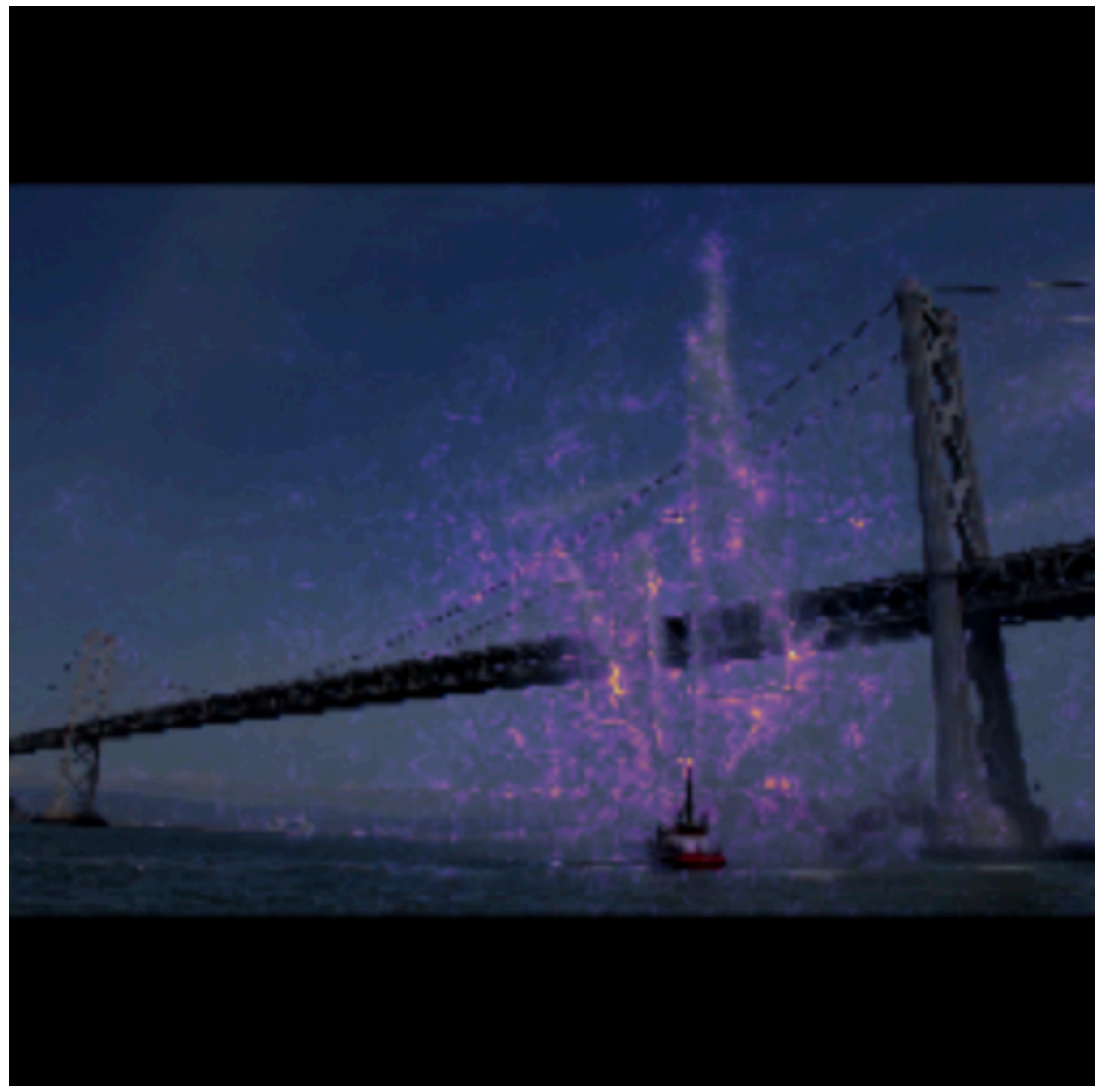
The Method of Integrated Gradients(IG)



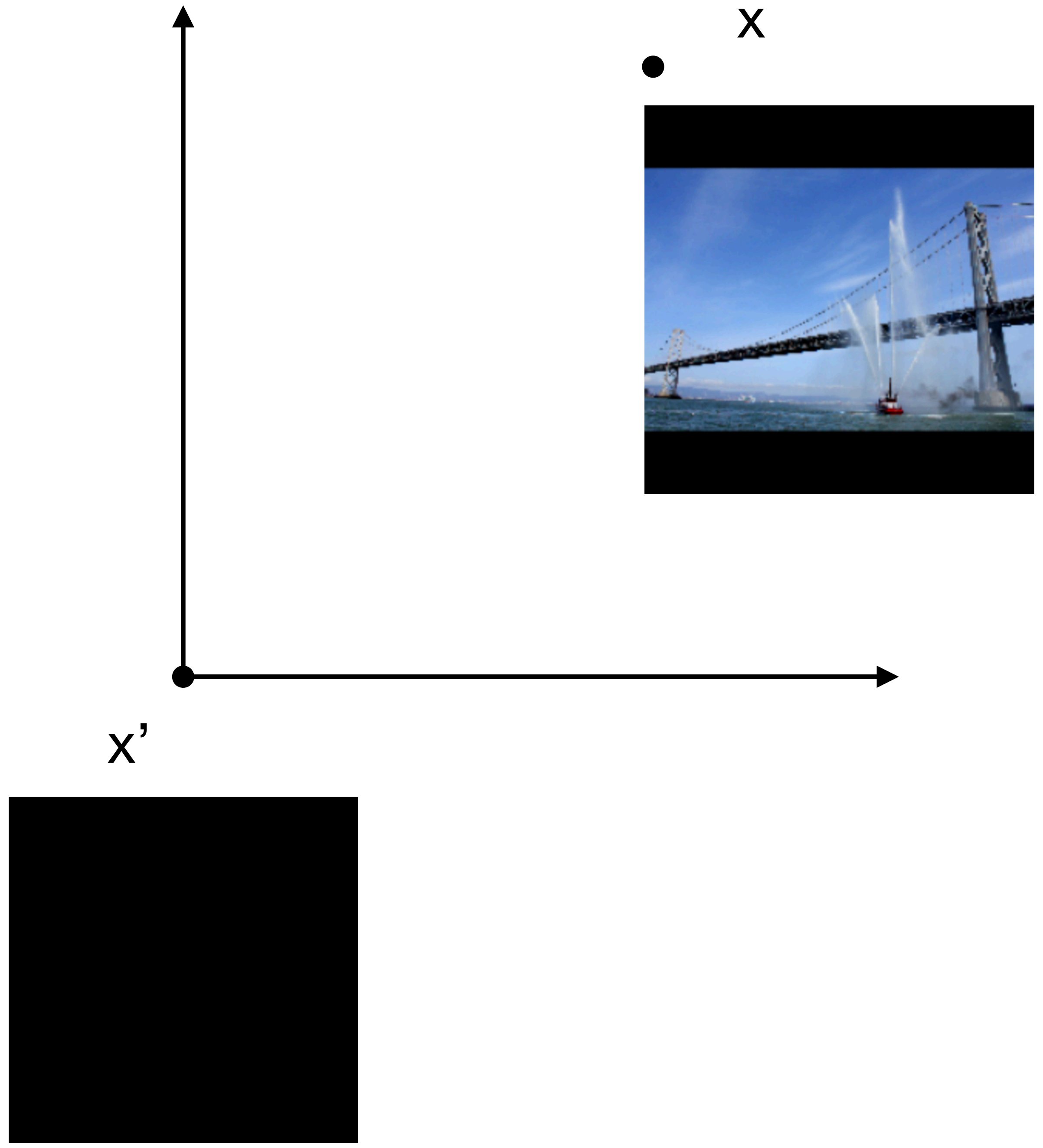
Fireboat & IG



The Method of Integrated Gradients(IG)

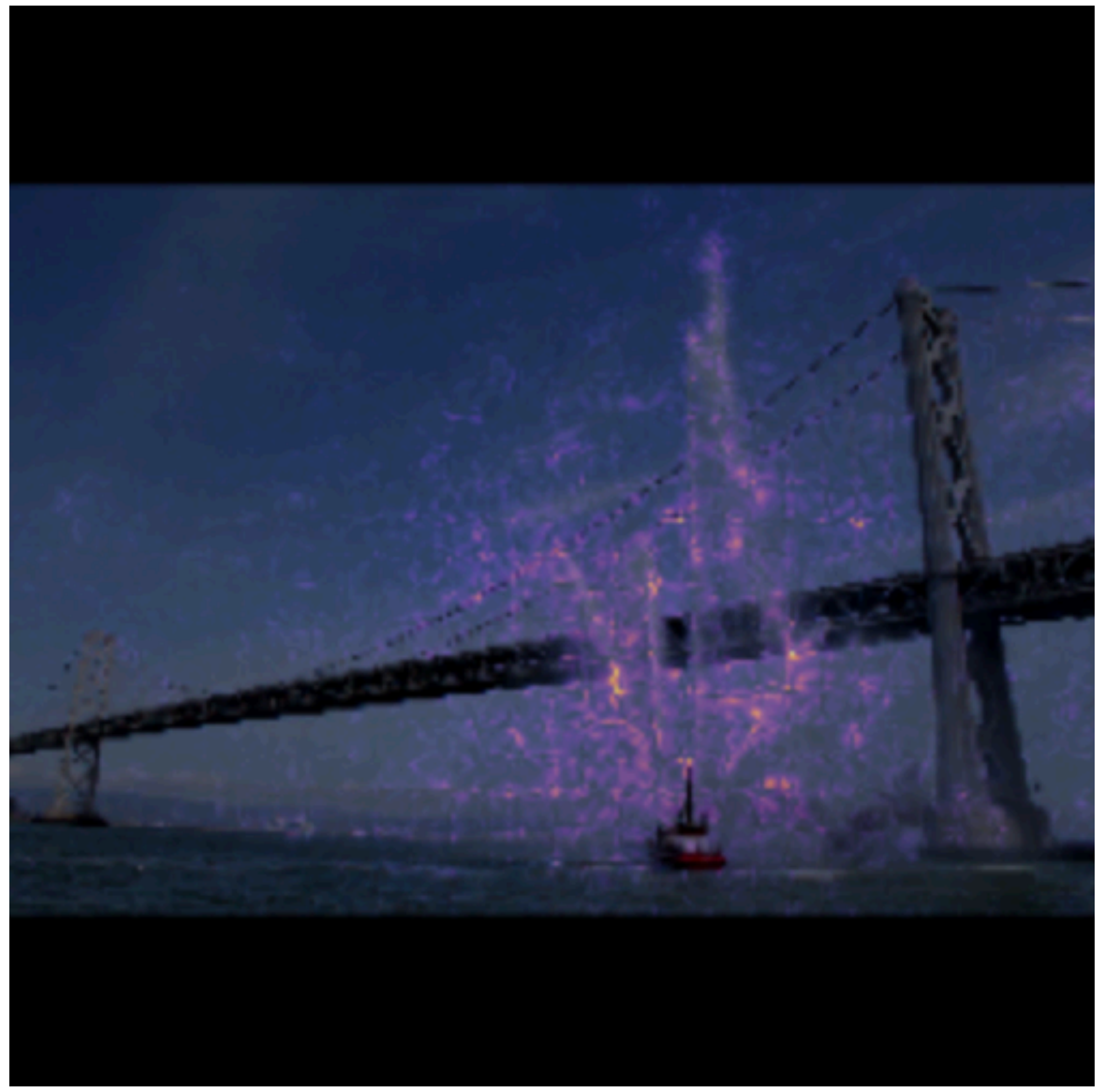


Fireboat & IG

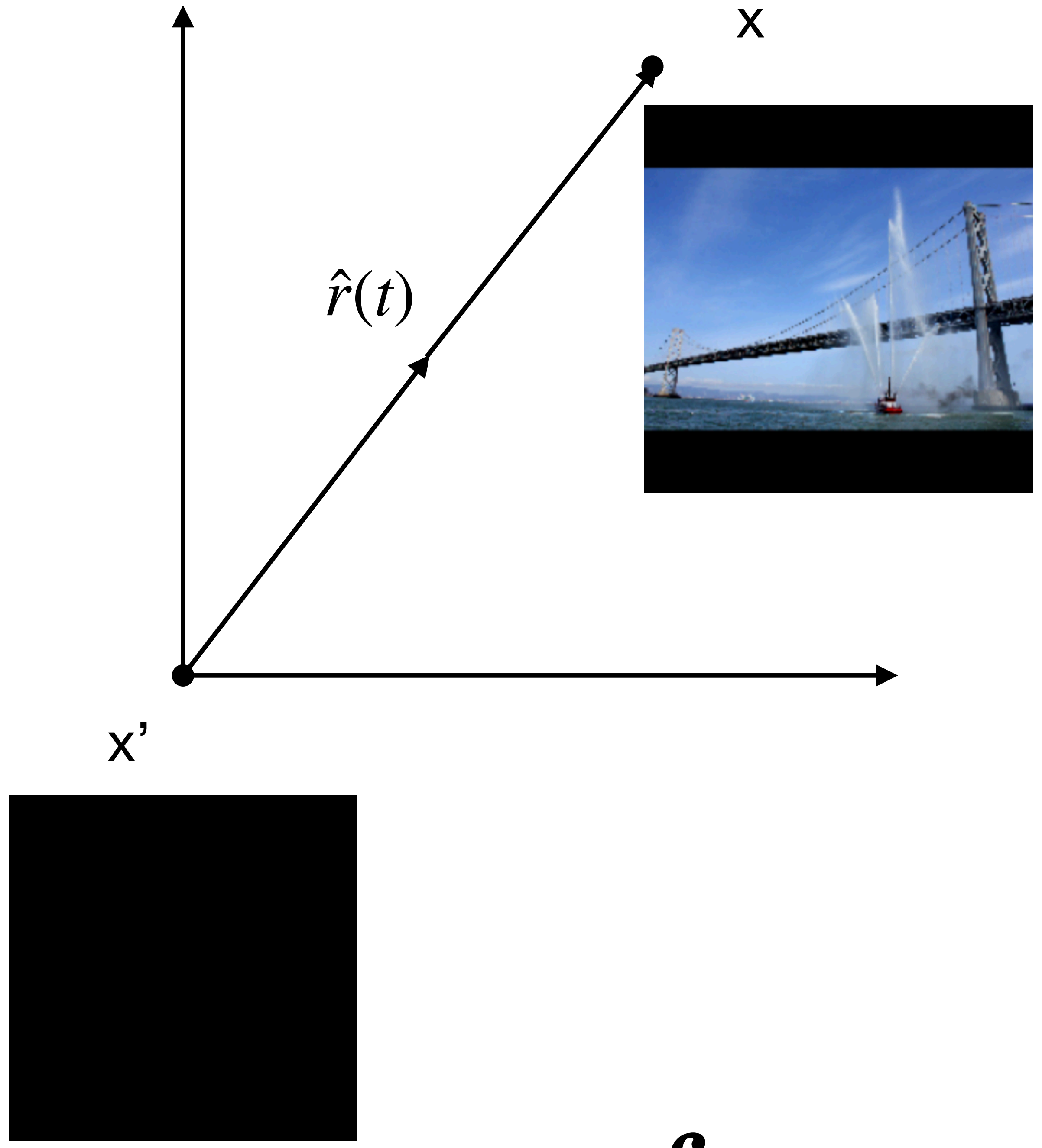


$$F(x) - F(x')$$

The Method of Integrated Gradients(IG)

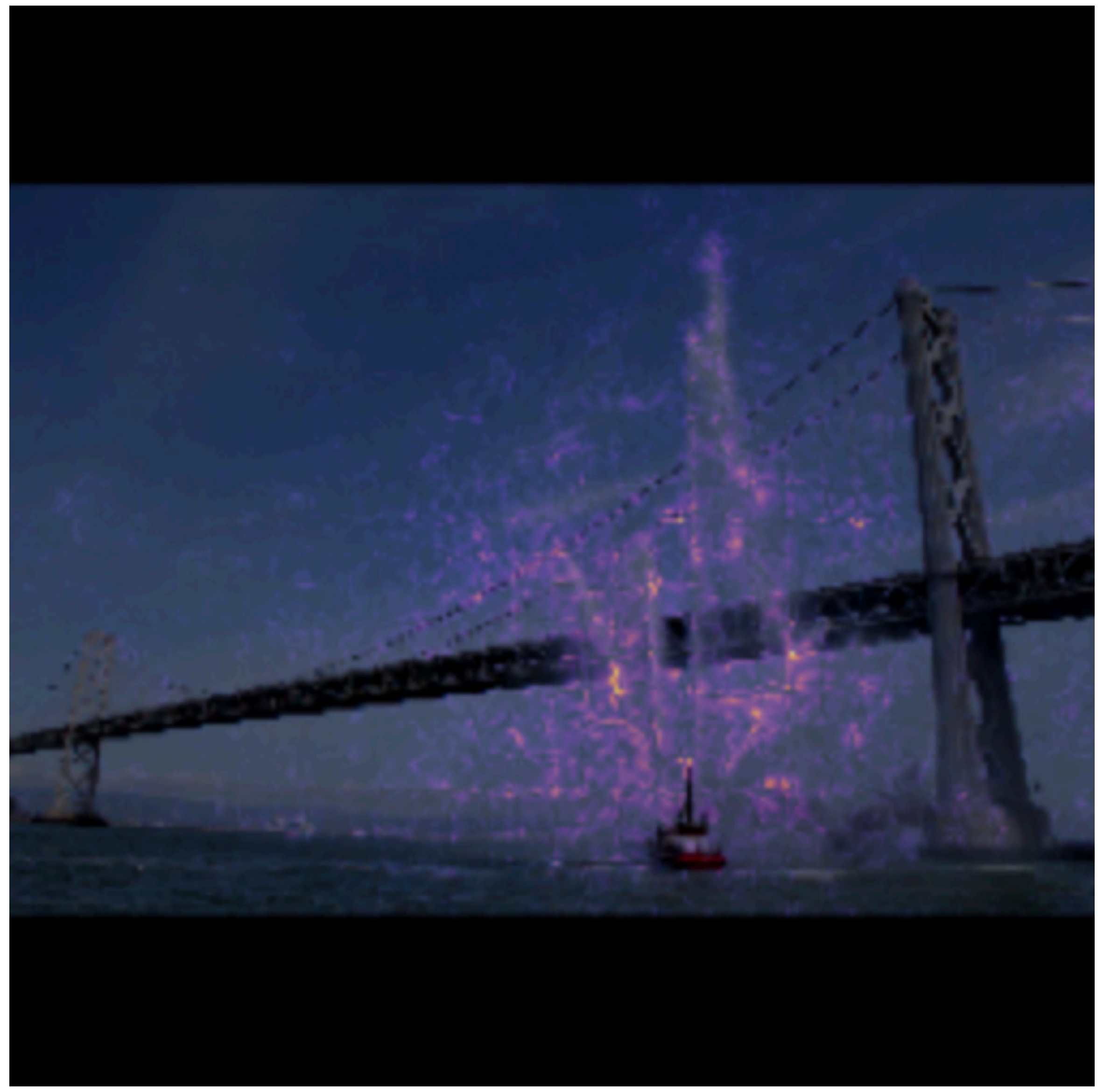


Fireboat & IG

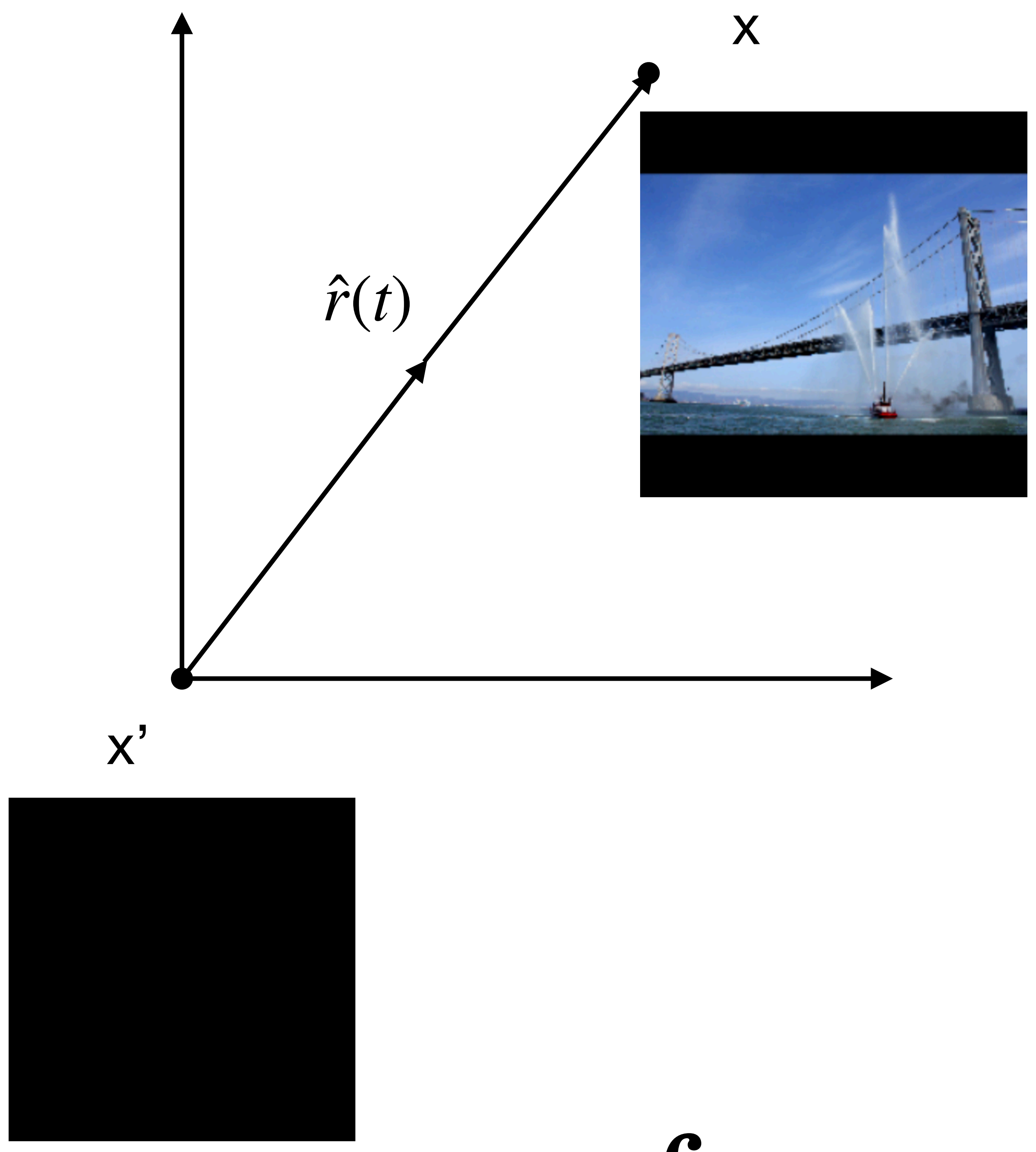


$$F(x) - F(x') = \int \nabla F \cdot d\hat{r}(t)$$

The Method of Integrated Gradients(IG)



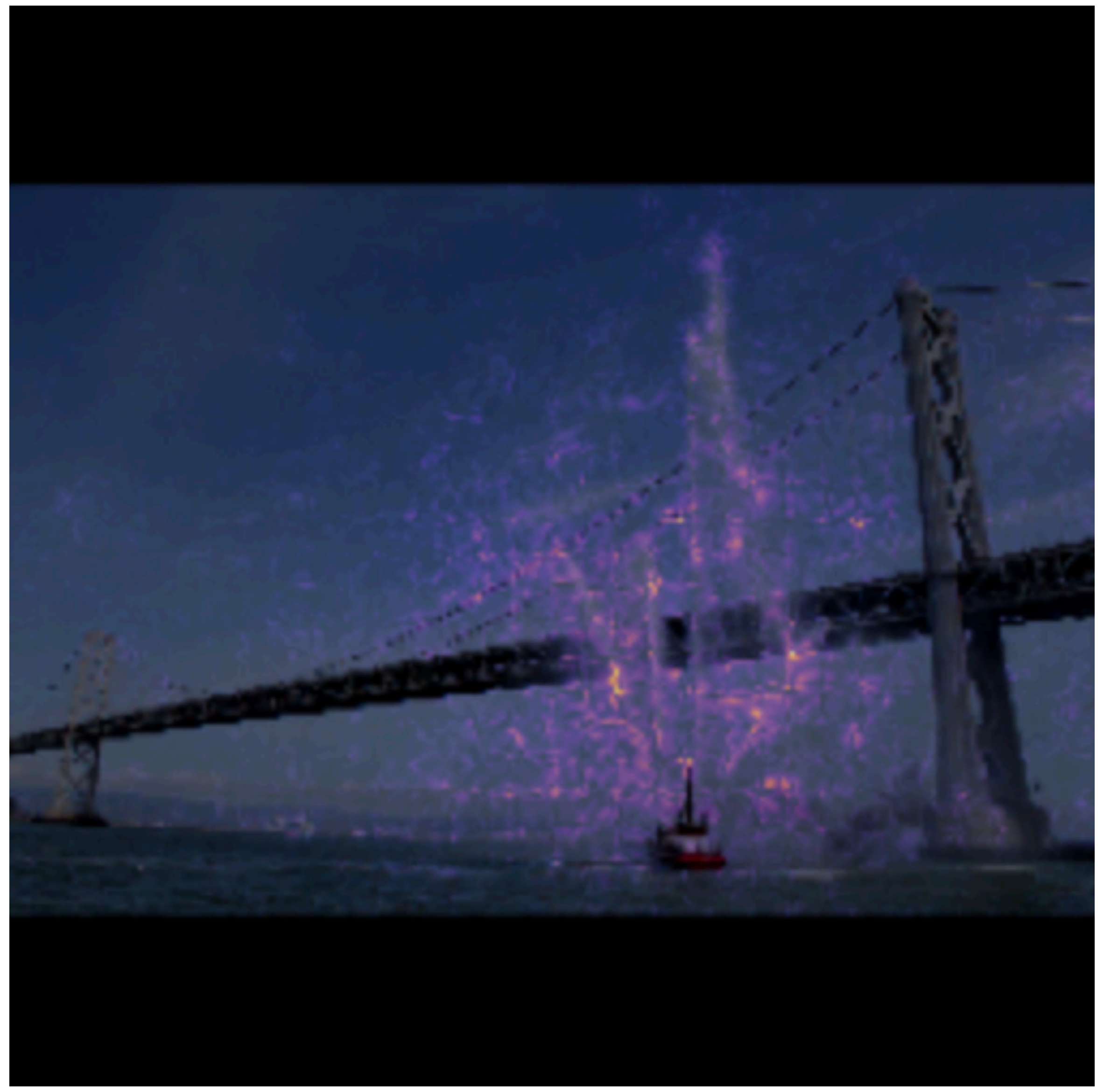
Fireboat & IG



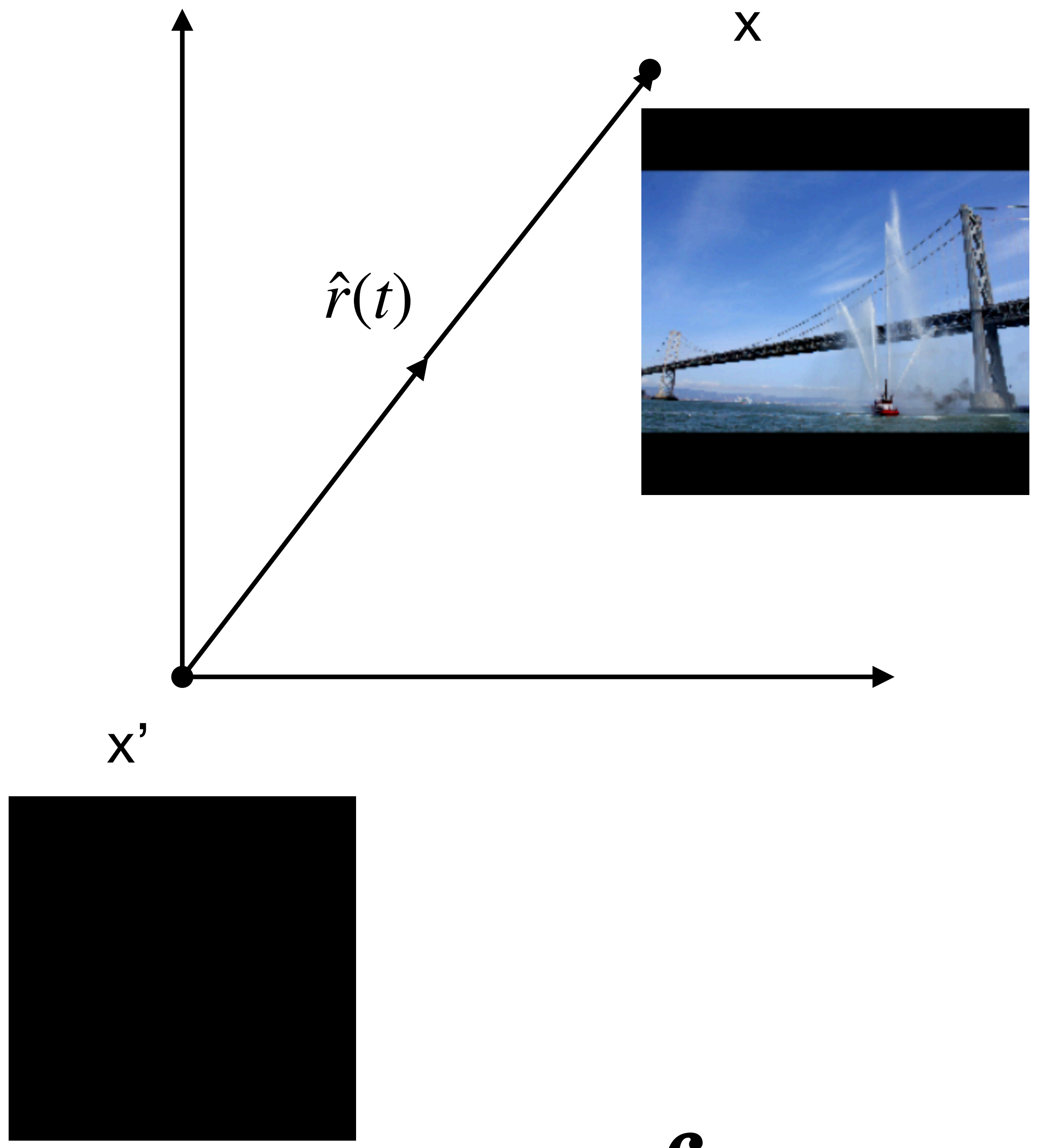
$$F(x) - F(x') = \int \nabla F \cdot d\hat{r}(t)$$

$$= \int \sum_{i=1}^{n} \frac{dF}{dr_i} \frac{dr_i}{dt} dt$$

The Method of Integrated Gradients(IG)



Fireboat & IG



$$F(x) - F(x') = \int \nabla F \cdot d\hat{r}(t)$$

$$= \int \sum_{i=1}^{n} \frac{dF}{dr_i} \frac{dr_i}{dt} dt$$

$$IG(x, x', F) = (\int \frac{dF}{dr_1} \frac{dr_1}{dt} dt, \ldots, \int \frac{dF}{dr_n} \frac{dr_n}{dt} dt)$$

Axiomatic Motivation

Axiomatic Motivation

- Sensitivity(a): If a change in one pixel causes a change in output, that pixel should have a non-zero attribution.

Axiomatic Motivation

- Sensitivity(a): If a change in one pixel causes a change in output, that pixel should have a non-zero attribution.
    - DeConv Nets [Zeiler & Fergus, 2014] and Guided Backprop [Springenberg et al., 2014] fail to satisfy.

Axiomatic Motivation

- Sensitivity(a): If a change in one pixel causes a change in output, that pixel should have a non-zero attribution.
    - DeConv Nets [Zeiler & Fergus, 2014] and Guided Backprop [Springenberg et al., 2014] fail to satisfy.
- Implementation Invariance: If two models are mathematically equivalent, they should receive equivalent attributions.

Axiomatic Motivation

- Sensitivity(a): If a change in one pixel causes a change in output, that pixel should have a non-zero attribution.
  - DeConv Nets [Zeiler & Fergus, 2014] and Guided Backprop [Springenberg et al., 2014] fail to satisfy.
- Implementation Invariance: If two models are mathematically equivalent, they should receive equivalent attributions.
  - Deeplift [Shrikumar er al, 2017] and LRP [Binder et al., 2016] fail to satisfy.

Axiomatic Motivation

- Sensitivity(a): If a change in one pixel causes a change in output, that pixel should have a non-zero attribution.
  - DeConv Nets [Zeiler & Fergus, 2014] and Guided Backprop [Springenberg et al., 2014] fail to satisfy.
- Implementation Invariance: If two models are mathematically equivalent, they should receive equivalent attributions.
  - Deeplift [Shrikumar er al, 2017] and LRP [Binder et al., 2016] fail to satisfy.
- IG satisfies these axioms, and claims to uniquely satisfy a group axioms.

Importing Game-Theoretic Results

Importing Game-Theoretic Results

- IG Uniqueness Strategy: Import results from game-theory (Aumann-Shaply)

Importing Game-Theoretic Results

- IG Uniqueness Strategy: Import results from game-theory (Aumann-Shaply)
- E.g., $A(x, x', F)$ game-theoretic attribution on model F, for input x, baseline x':

Importing Game-Theoretic Results

- IG Uniqueness Strategy: Import results from game-theory (Aumann-Shaply)
- E.g., $A(x, x', F)$ game-theoretic attribution on model F, for input x, baseline x':
    - Dummy: If $\partial_i F \equiv 0$, then $A_i(x, x', F) = 0$

Importing Game-Theoretic Results

- IG Uniqueness Strategy: Import results from game-theory (Aumann-Shaply)
- E.g., $A(x, x', F)$ game-theoretic attribution on model F, for input x, baseline x':
    - Dummy: If $\partial_i F \equiv 0$, then $A_i(x, x', F) = 0$
    - Linearity: $A(x, x', F + G) = A(x, x', F) + A(x, x', G)$

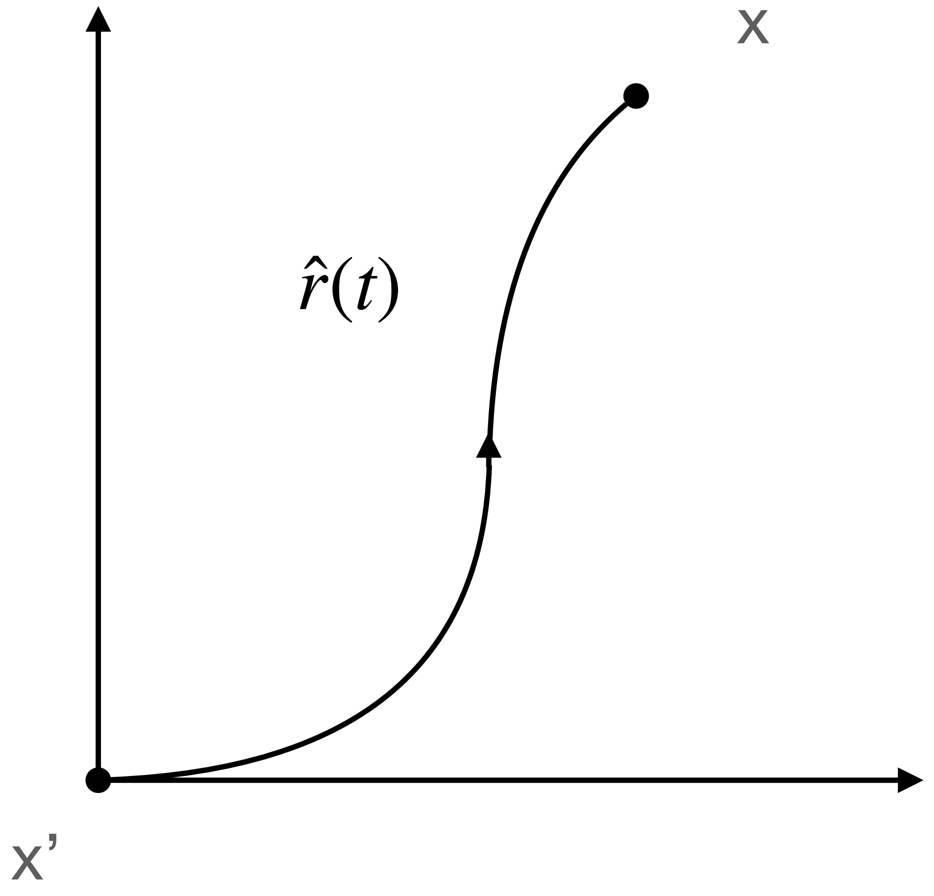
Importing Game-Theoretic Results

- IG Uniqueness Strategy: Import results from game-theory (Aumann-Shaply)
- E.g., $A(x, x', F)$ game-theoretic attribution on model F, for input x, baseline x':
  - Dummy: If $\partial_i F \equiv 0$, then $A_i(x, x', F) = 0$
  - Linearity: $A(x, x', F + G) = A(x, x', F) + A(x, x', G)$
  - Completeness: $\sum_{i=1}^{n} A_i(x, x', F) = F(x) - F(x')$

Importing Game-Theoretic Results

- IG Uniqueness Strategy: Import results from game-theory (Aumann-Shaply)
- E.g., $A(x, x', F)$ game-theoretic attribution on model F, for input x, baseline x':
  - Dummy: If $\partial_i F \equiv 0$, then $A_i(x, x', F) = 0$
  - Linearity: $A(x, x', F + G) = A(x, x', F) + A(x, x', G)$
  - Completeness: $\sum_{i=1}^{n} A_i(x, x', F) = F(x) - F(x')$

[Friedman, 2004]



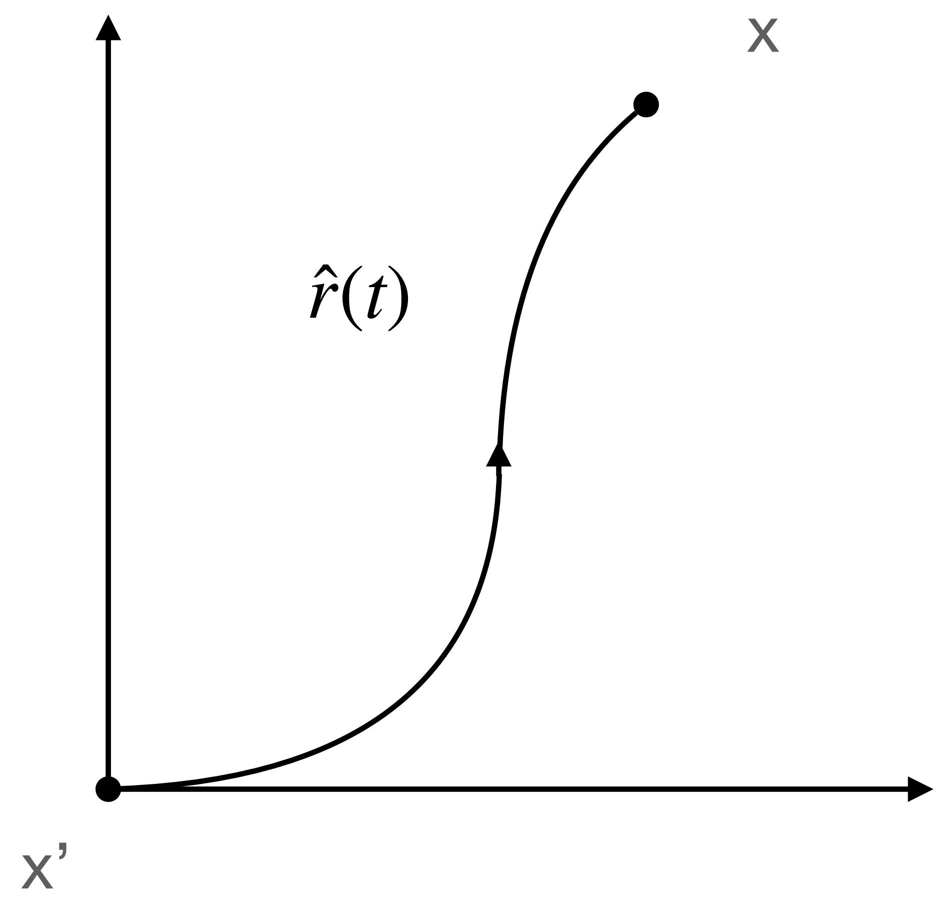
A is an accumulation of gradients for some monotone path integral.

$$A(x, x', F) = (\int_0^1 \frac{dF}{dr_1}\frac{dr_1}{dt}dt, \ldots, \int_0^1 \frac{dF}{dr_n}\frac{dr_n}{dt}dt)$$

Importing Game-Theoretic Results

- IG Uniqueness Strategy: Import results from game-theory (Aumann-Shaply)
- E.g., $A(x, x', F)$ game-theoretic attribution on model F, for input x, baseline x':
  - Dummy: If $\partial_i F \equiv 0$, then $A_i(x, x', F) = 0$
  - Linearity: $A(x, x', F + G) = A(x, x', F) + A(x, x', G)$
  - Completeness: $\sum_{i=1}^{n} A_i(x, x', F) = F(x) - F(x')$
- <u>IG claim</u> - <u>the deep learning analogue holds also.</u>

[Friedman, 2004]



A is an accumulation of gradients for some monotone path integral.

$$A(x, x', F) = (\int_0^1 \frac{dF}{dr_1}\frac{dr_1}{dt}dt, \ldots, \int_0^1 \frac{dF}{dr_n}\frac{dr_n}{dt}dt)$$

Classical Game Theory vs DL Context

Property	Classical Game Theory [Friedman, 2004]	Object Classification
Attribute Restrictions	$A \geq 0$	$A \in \mathbb{R}^n$
Monotonic Model?	Yes	No
Model Smoothness	$F \in C^1$	$F \in C^0$
Baseline	$x' = 0$	$x' \in [0,1]^n$

Classical Game Theory vs DL Context

Property	Classical Game Theory [Friedman, 2004]	Object Classification
Attribute Restrictions	$A \geq 0$	$A \in \mathbb{R}^n$
Monotonic Model?	Yes	No
Model Smoothness	$F \in C^1$	$F \in C^0$
Baseline	$x' = 0$	$x' \in [0,1]^n$

- The IG uniqueness claim does not hold.

Classical Game Theory vs DL Context

Property	Classical Game Theory [Friedman, 2004]	Object Classification
Attribute Restrictions	$A \geq 0$	$A \in \mathbb{R}^n$
Monotonic Model?	Yes	No
Model Smoothness	$F \in C^1$	$F \in C^0$
Baseline	$x' = 0$	$x' \in [0,1]^n$

- The IG uniqueness claim does not hold.
- We establish uniqueness claims with an additional axiom.

Targeting Regions for Neuron Attributions

Targeting Regions for Neuron Attributions

- IG - Which inputs contributed to an output?

Targeting Regions for Neuron Attributions

- IG - Which inputs contributed to an output?
- Neuron IG - Which neurons contributed to an output? [Dhamdhere et al., 2018]

Targeting Regions for Neuron Attributions

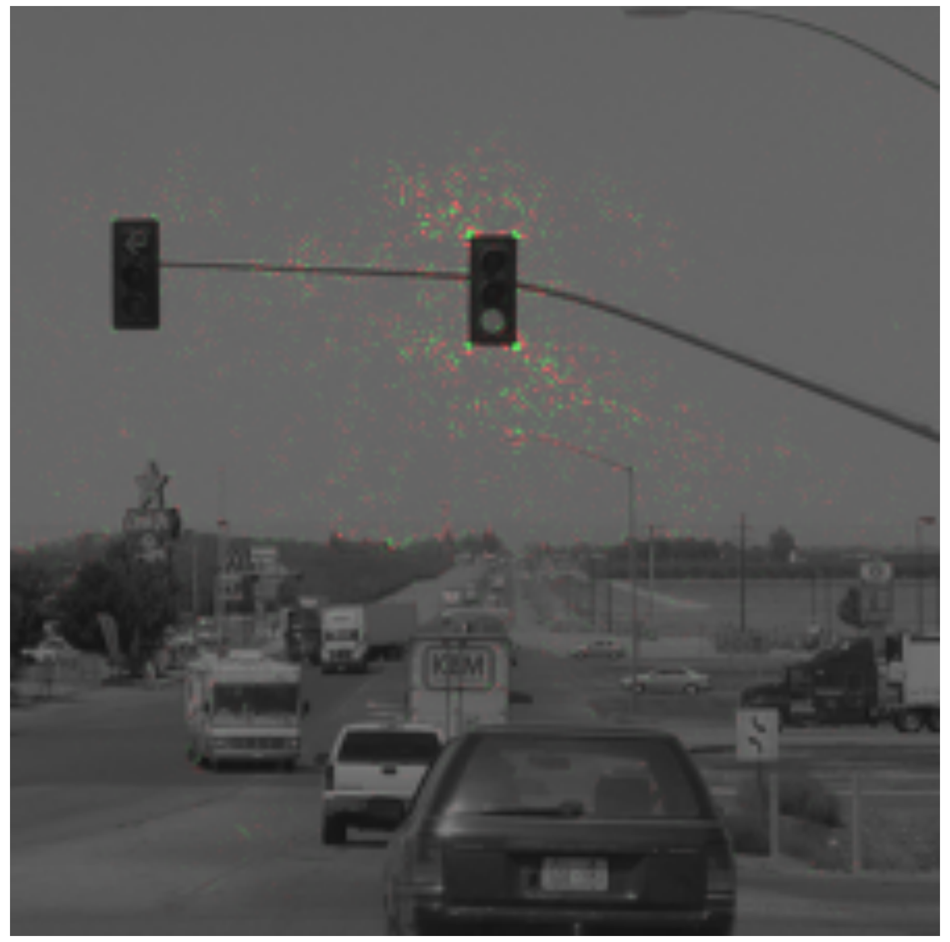
- IG - Which inputs contributed to an output?
- Neuron IG - Which neurons contributed to an output? [Dhamdhere et al., 2018]

Image of Stoplights

Targeting Regions for Neuron Attributions

- IG - Which inputs contributed to an output?
- Neuron IG - Which neurons contributed to an output? [Dhamdhere et al., 2018]

Image of Stoplights

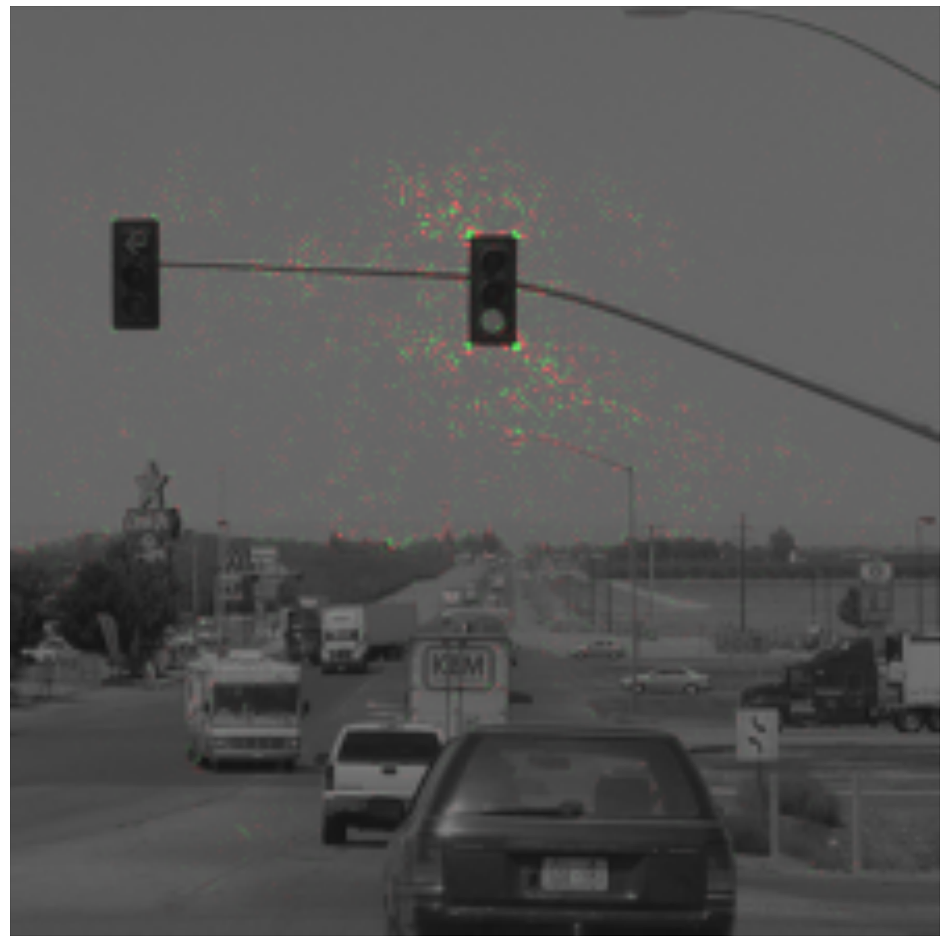


IG Attribution to Inputs

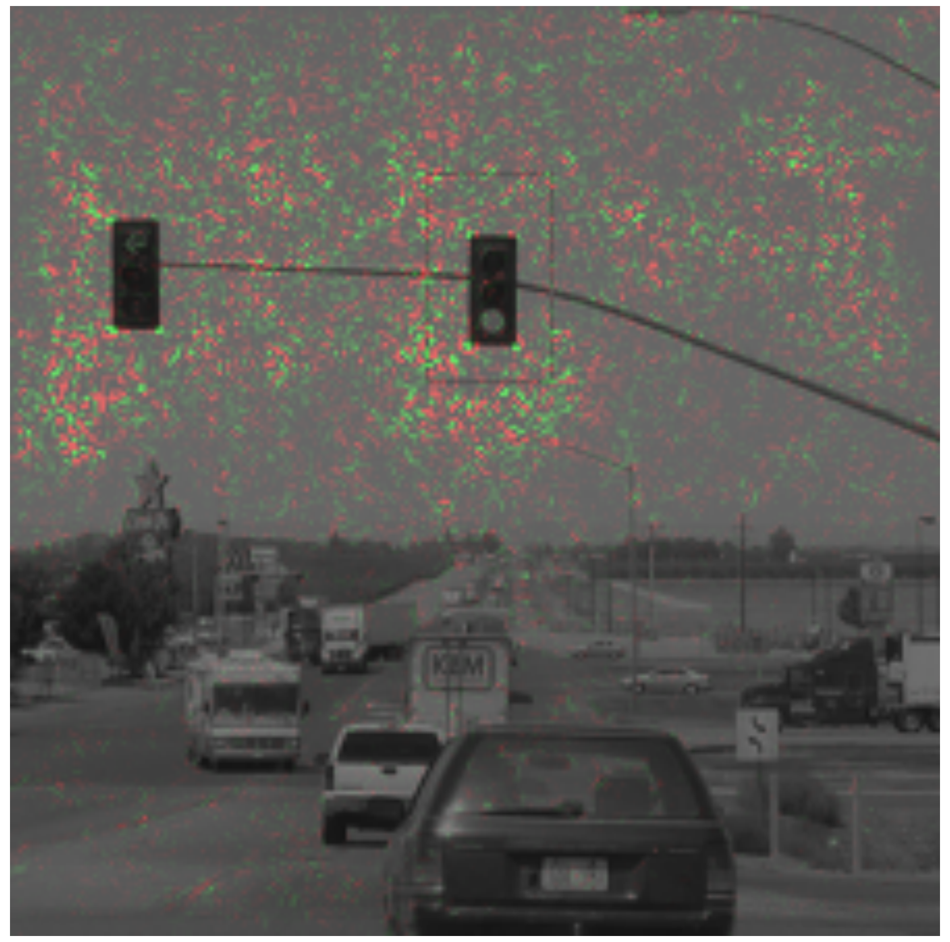
Targeting Regions for Neuron Attributions

- IG - Which inputs contributed to an output?
- Neuron IG - Which neurons contributed to an output? [Dhamdhere et al., 2018]

Image of Stoplights



IG Attribution to Inputs

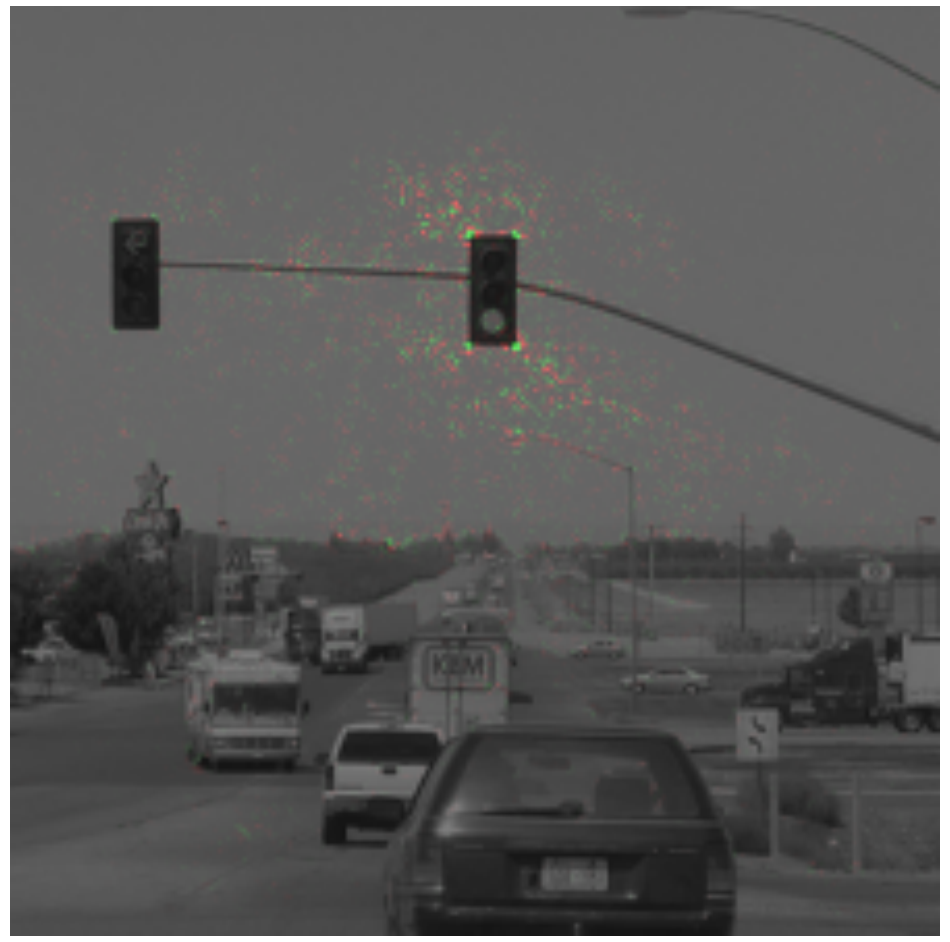


IG after Top-1% Pruned

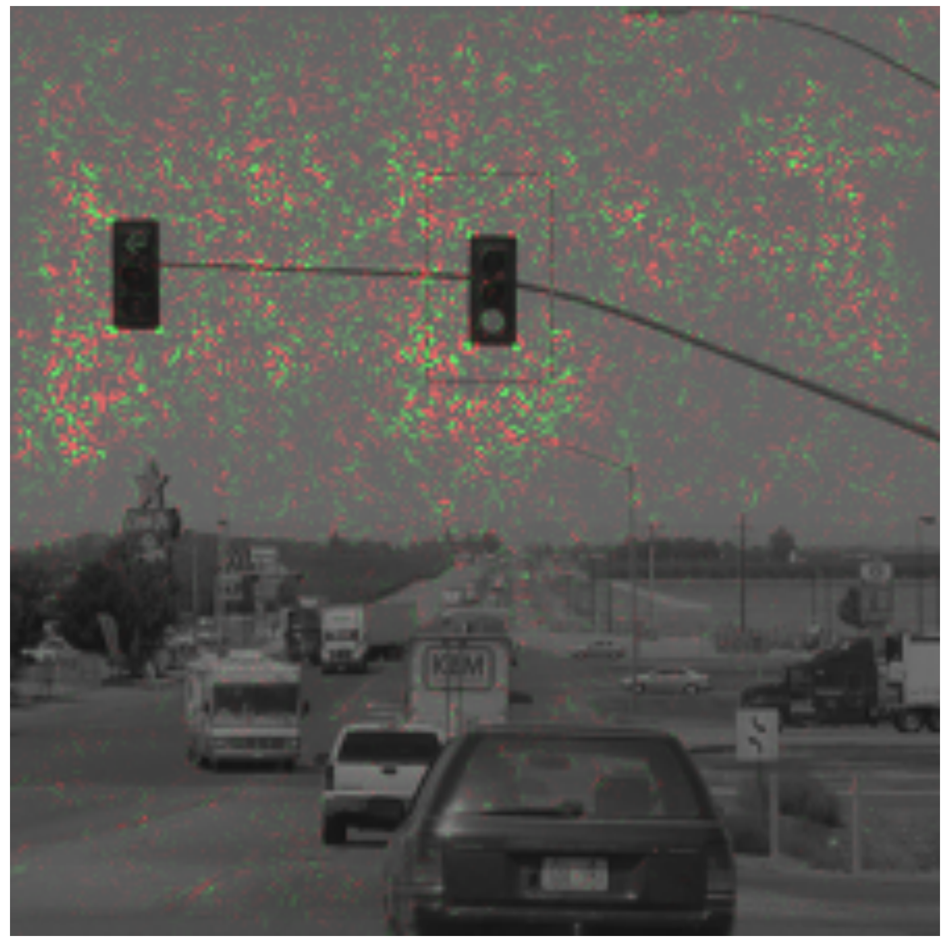
Targeting Regions for Neuron Attributions

- IG - Which inputs contributed to an output?
- Neuron IG - Which neurons contributed to an output? [Dhamdhere et al., 2018]
- Targeted Neuron IG - Which neurons associated with the targeted region contributed to the output?

Image of Stoplights



IG Attribution to Inputs

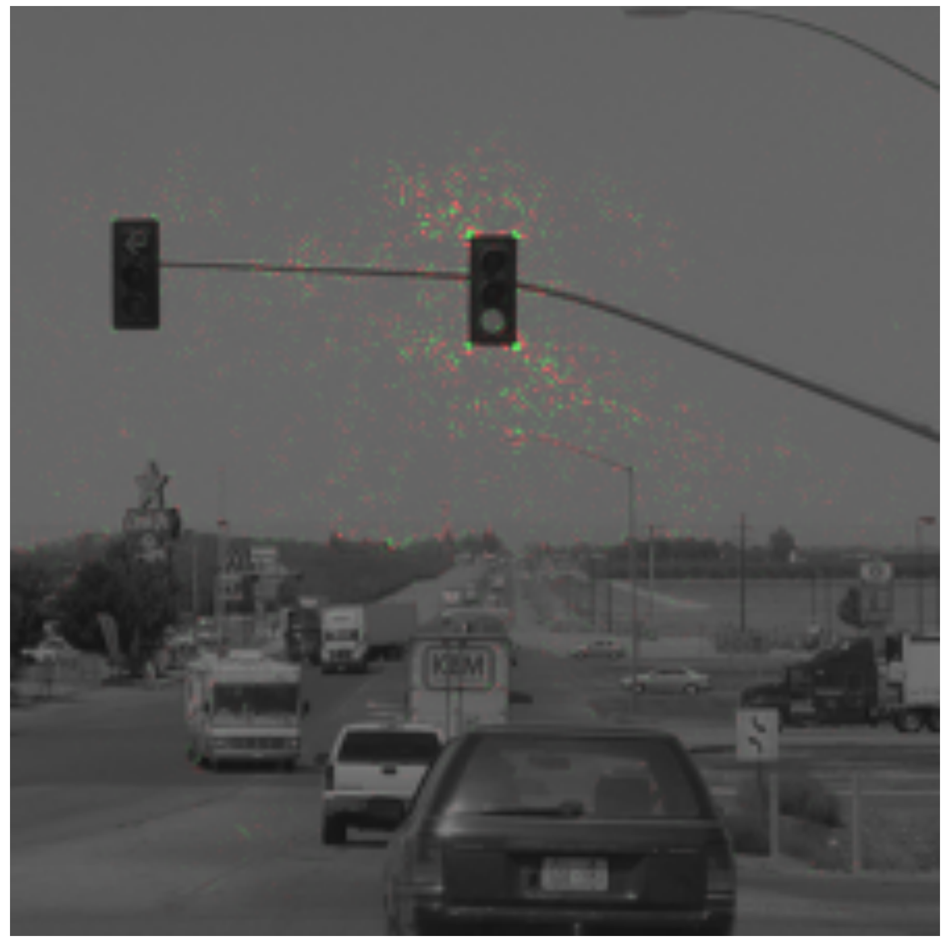


IG after Top-1% Pruned

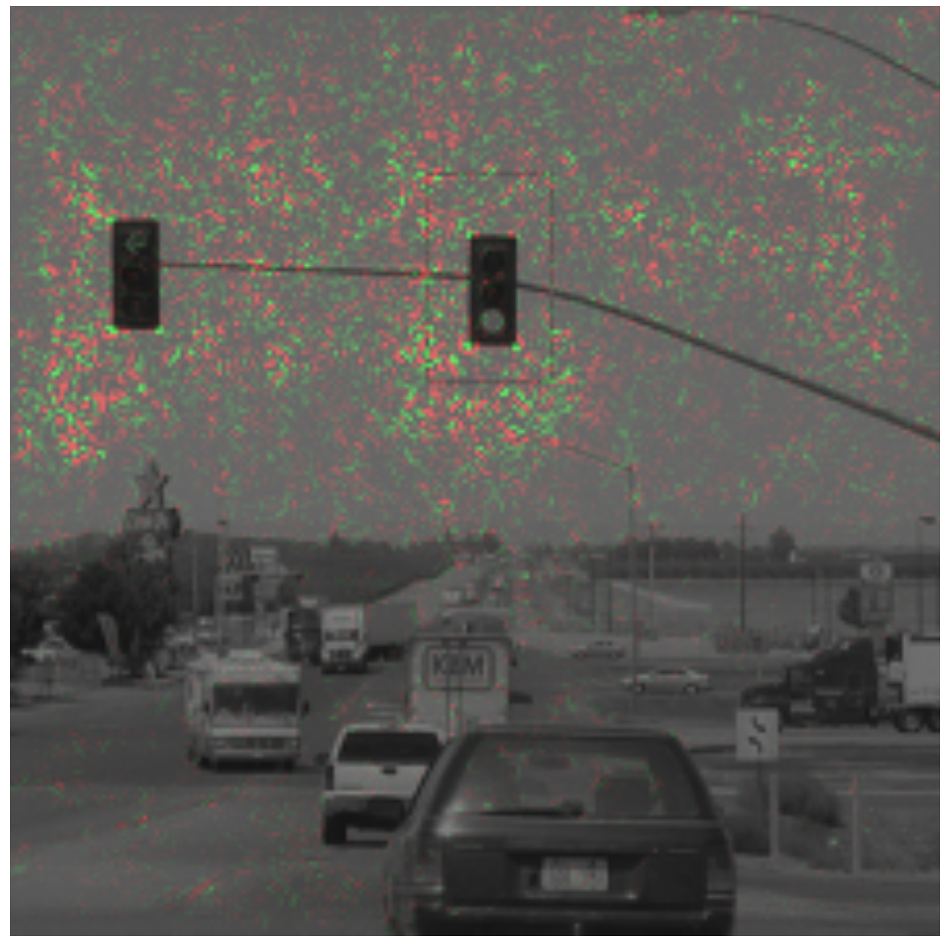
Targeting Regions for Neuron Attributions

- IG - Which inputs contributed to an output?
- Neuron IG - Which neurons contributed to an output? [Dhamdhere et al., 2018]
- Targeted Neuron IG - Which neurons associated with the targeted region contributed to the output?

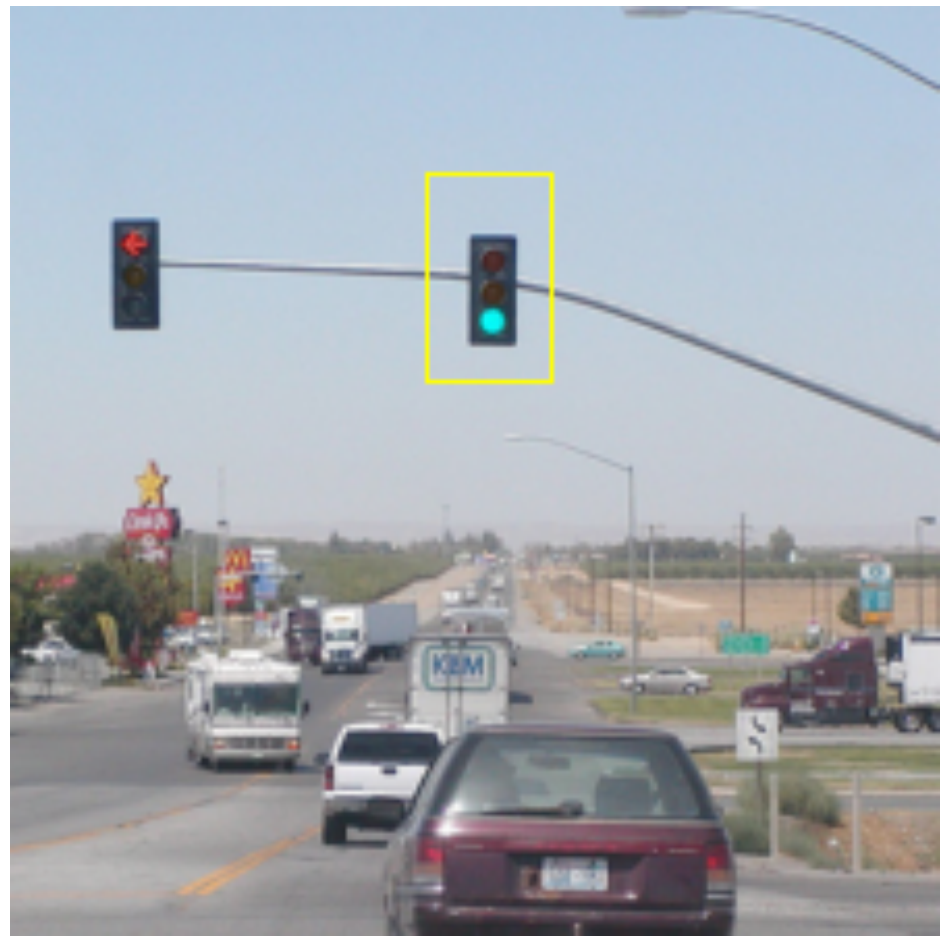
Image of Stoplights



IG Attribution to Inputs



IG after Top-1% Pruned



Targeting Regions for Neuron Attributions

- IG - Which inputs contributed to an output?
- Neuron IG - Which neurons contributed to an output? [Dhamdhere et al., 2018]
- Targeted Neuron IG - Which neurons associated with the targeted region contributed to the output?

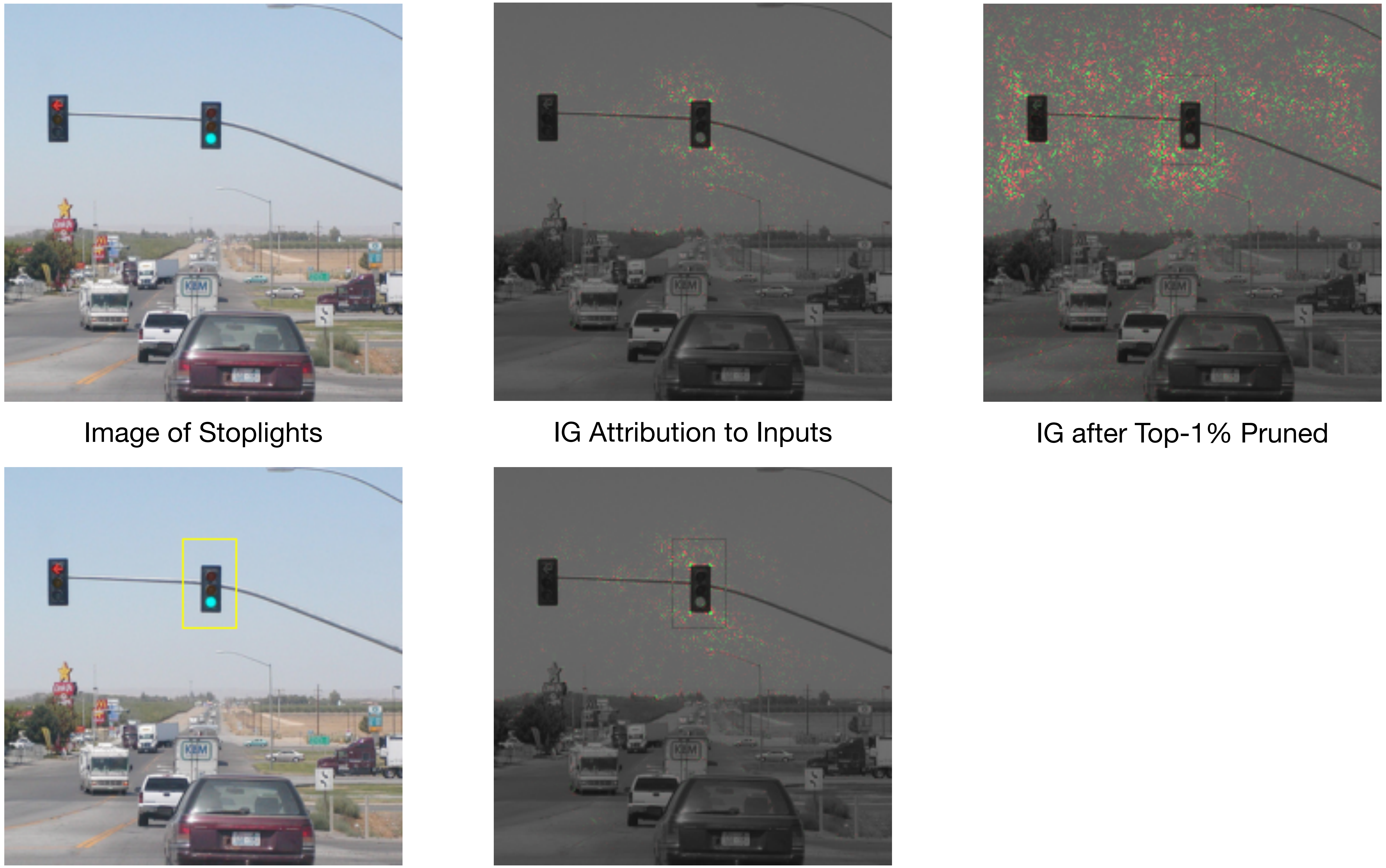


Image of Stoplights

IG Attribution to Inputs

IG after Top-1% Pruned

Targeting Regions for Neuron Attributions

- IG - Which inputs contributed to an output?
- Neuron IG - Which neurons contributed to an output? [Dhamdhere et al., 2018]
- Targeted Neuron IG - Which neurons associated with the targeted region contributed to the output?

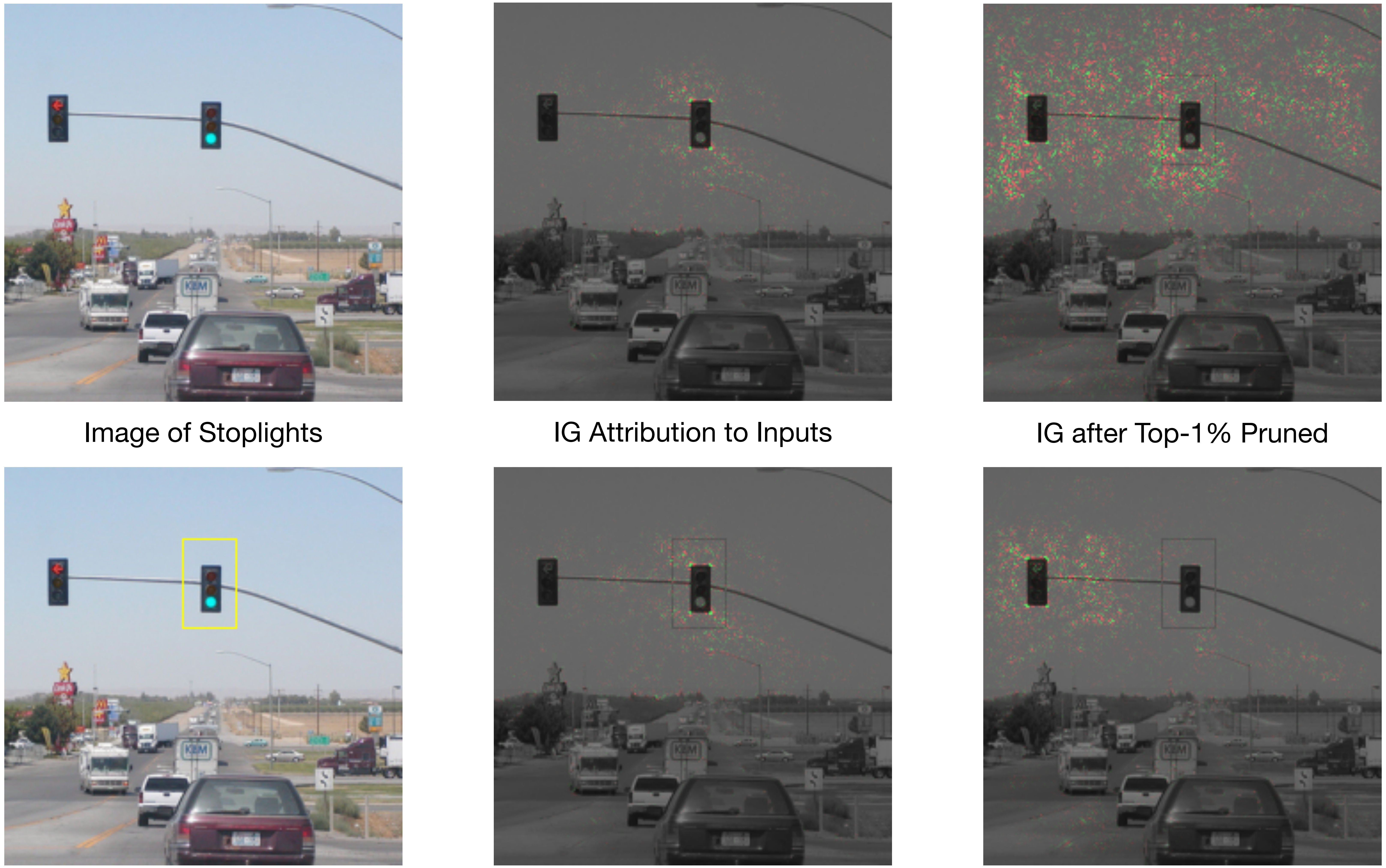


Image of Stoplights

IG Attribution to Inputs

IG after Top-1% Pruned