— DAVE —

Distribution-Aware **A**ttribution via **V**iT Gradient D**E**composition

Adam Wróbel *

Siddhartha Gairola *

Jacek Tabor

Bernt Schiele

Bartosz Zieliński

Dawid Rymarczyk

JAGIELLONIAN UNIVERSITY IN KRAKÓW

MAX PLANCK INSTITUTE FOR INFORMATICS

ICML International Conference On Machine Learning

ViT Attribution Gap

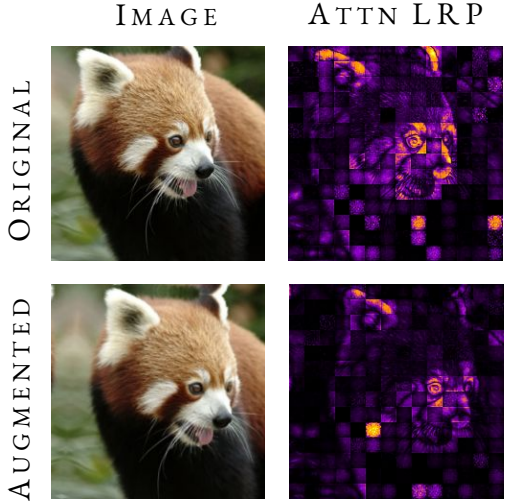
Architecture-Induced Artifacts

- **ViTs** dominate modern vision.
- **Pixel-level** attributions suffer from **artifacts**.
- **Patch-level** attributions **lack precision**.
- Stable high-resolution explanations are missing.
- **DAVE: artifact-free**, **pixel-level** attribution.

ViT Attributions Gap

Architecture-Induced Artifacts

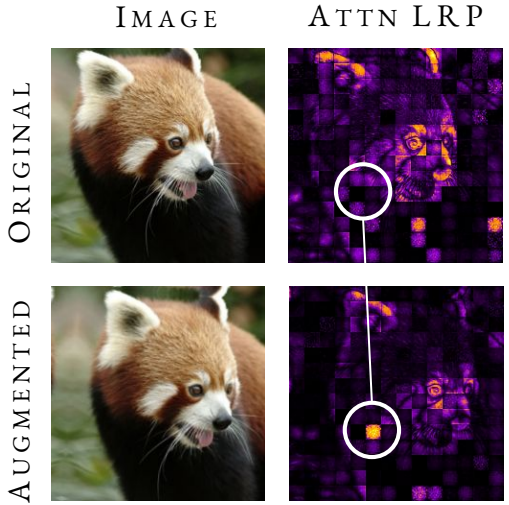
- **ViTs** dominate modern vision.
- **Pixel-level** attributions suffer from **artifacts**.
- **Patch-level** attributions **lack precision**.
- Stable high-resolution explanations are missing.
- **DAVE: artifact-free, pixel-level** attribution.



ViT Attributions Gap

Architecture-Induced Artifacts

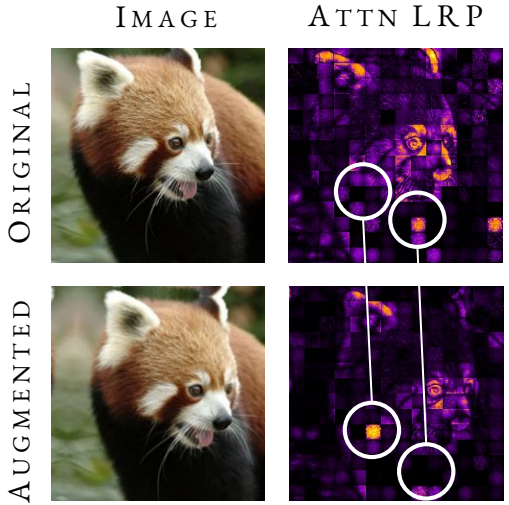
- **ViTs** dominate modern vision.
- **Pixel-level** attributions suffer from **artifacts**.
- **Patch-level** attributions **lack precision**.
- Stable high-resolution explanations are missing.
- **DAVE: artifact-free, pixel-level** attribution.



ViT Attributions Gap

Architecture-Induced Artifacts

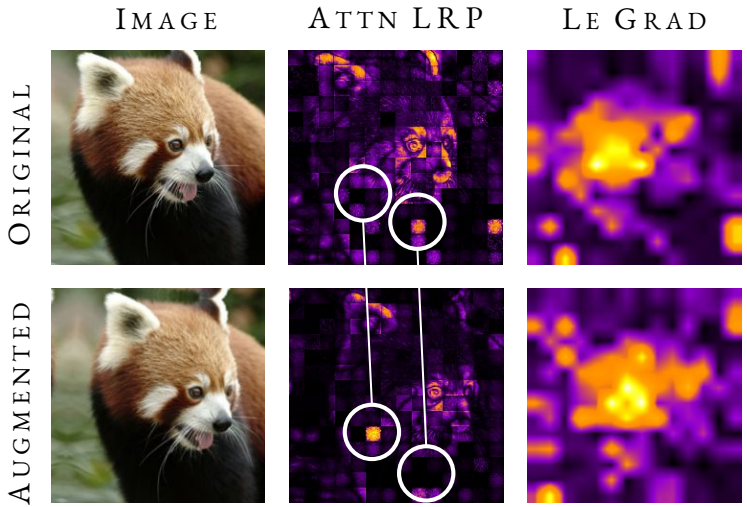
- **ViTs** dominate modern vision.
- **Pixel-level** attributions suffer from **artifacts**.
- **Patch-level** attributions **lack precision**.
- Stable high-resolution explanations are missing.
- **DAVE: artifact-free, pixel-level** attribution.



ViT Attributions Gap

Architecture-Induced Artifacts

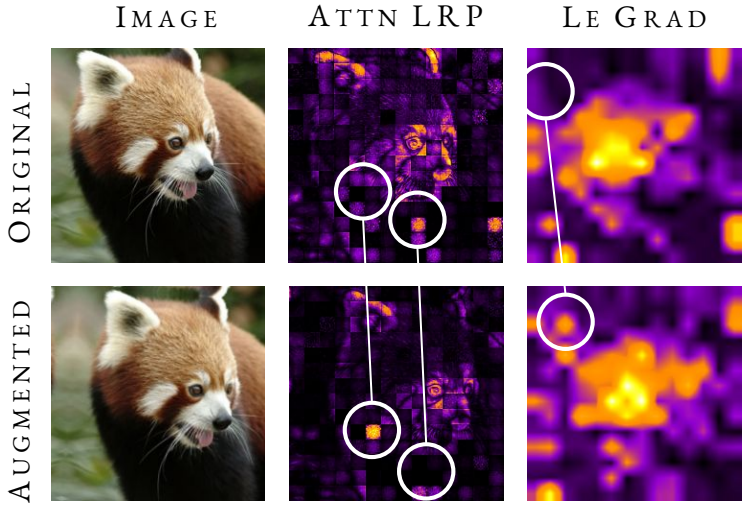
- **ViTs** dominate modern vision.
- **Pixel-level** attributions suffer from **artifacts**.
- **Patch-level** attributions **lack precision**.
- Stable high-resolution explanations are missing.
- **DAVE: artifact-free, pixel-level** attribution.



ViT Attributions Gap

Architecture-Induced Artifacts

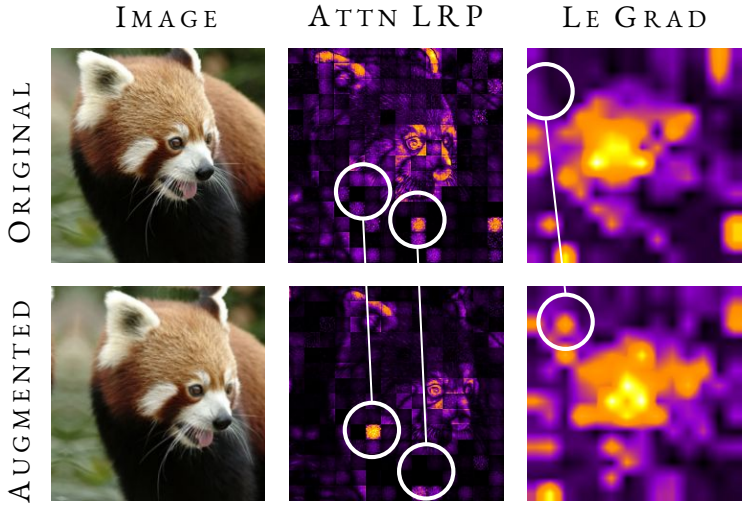
- **ViTs** dominate modern vision.
- **Pixel-level** attributions suffer from **artifacts**.
- **Patch-level** attributions **lack precision**.
- Stable high-resolution explanations are missing.
- **DAVE: artifact-free, pixel-level** attribution.



ViT Attributions Gap

Architecture-Induced Artifacts

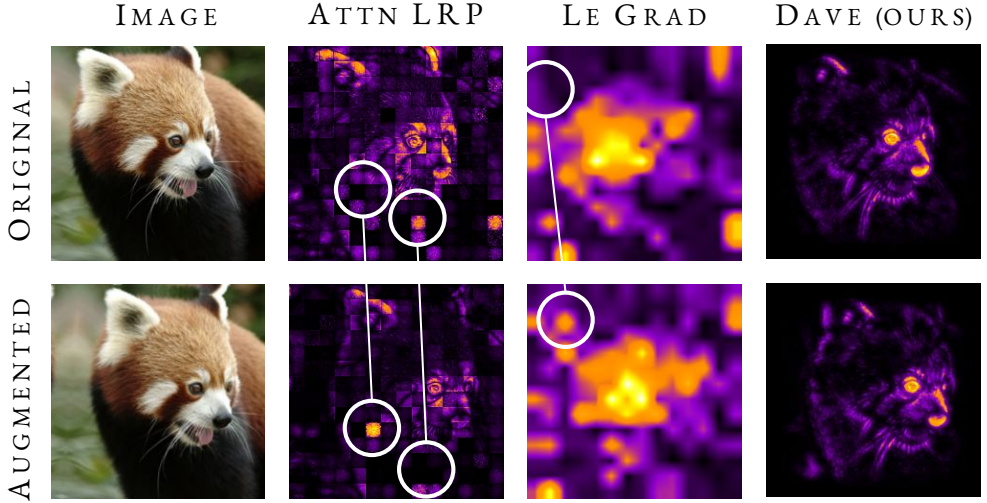
- **ViTs** dominate modern vision.
- **Pixel-level** attributions suffer from **artifacts**.
- **Patch-level** attributions **lack precision**.
- Stable high-resolution explanations are missing.
- **DAVE: artifact-free, pixel-level** attribution.



ViT Attributions Gap

Architecture-Induced Artifacts

- **ViTs** dominate modern vision.
- **Pixel-level** attributions suffer from **artifacts**.
- **Patch-level** attributions **lack precision**.
- Stable high-resolution explanations are missing.
- **DAVE: artifact-free, pixel-level** attribution.

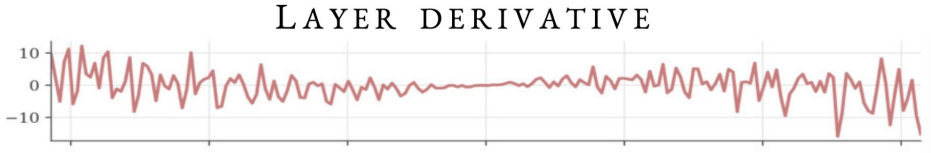


DAVE Construction

Step I: Effective Transformation

$D_x F$

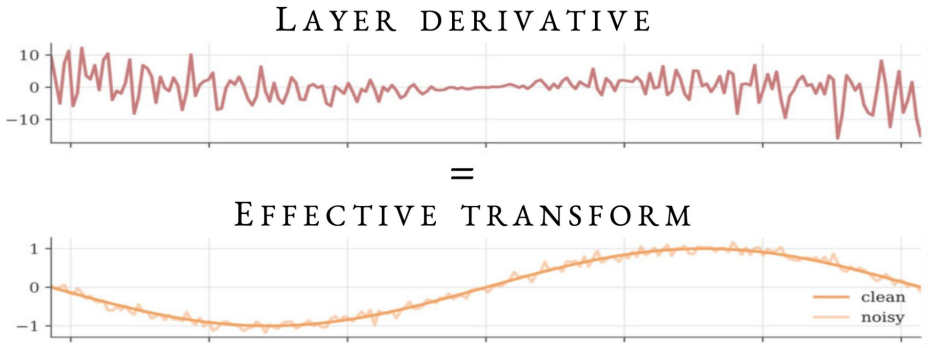
layer derivative



DAVE Construction

Step I: Effective Transformation

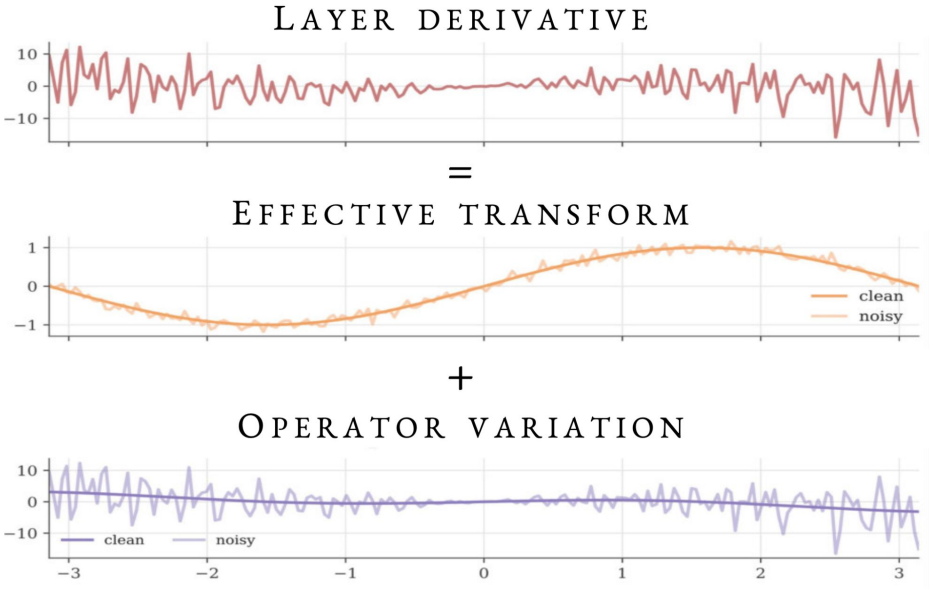
$$\underbrace{D_{\boldsymbol{X}} F}_{\text{layer derivative}} = \underbrace{L(\boldsymbol{X})}_{\text{effective transformation}}$$



DAVE Construction

Step I: Effective Transformation

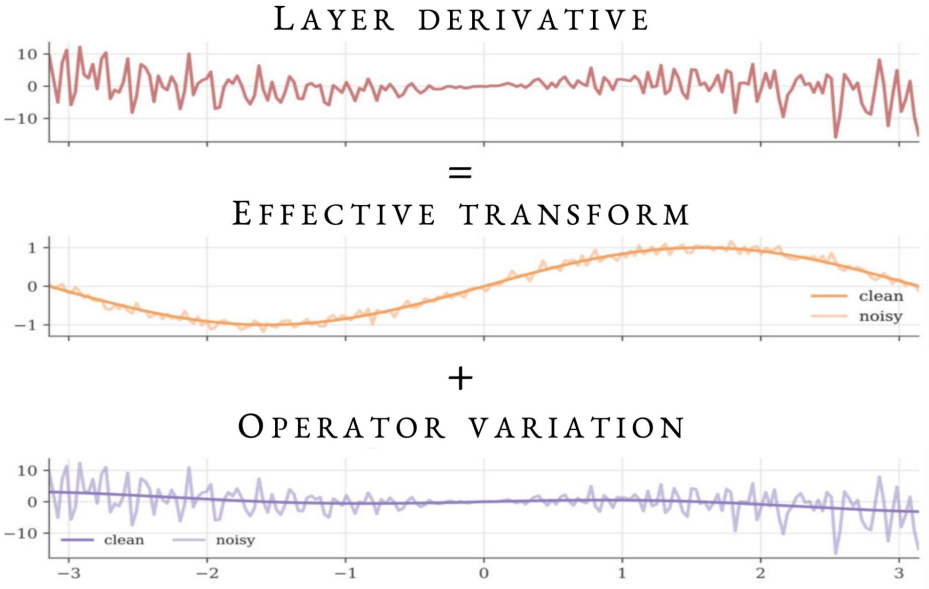
$$\underbrace{D_{\boldsymbol{X}} F}_{\text{layer derivative}} = \underbrace{L(\boldsymbol{X})}_{\text{effective transformation}} + \underbrace{((D_{\boldsymbol{X}} L(\boldsymbol{X})(\cdot))\,\boldsymbol{X})}_{\text{operator variation}}$$



DAVE Construction

Step I: Effective Transformation

$$\underbrace{D_{\boldsymbol{X}}F}_{\text{layer derivative}} = \underbrace{L(\boldsymbol{X})}_{\text{effective transformation}} + \underbrace{((D_{\boldsymbol{X}}L(\boldsymbol{X})(\cdot))\,\boldsymbol{X})}_{\text{operator variation}}$$



DAVE Construction

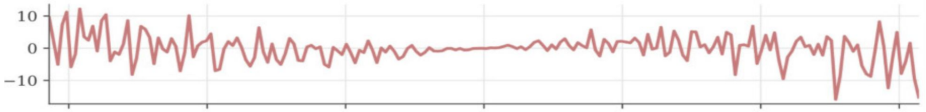
Step I: Effective Transformation

$$D_{\boldsymbol{X}}F = L(\boldsymbol{X}) + ((D_{\boldsymbol{X}}L(\boldsymbol{X})(\cdot))\,\boldsymbol{X})$$

layer derivative = effective transformation + operator variation

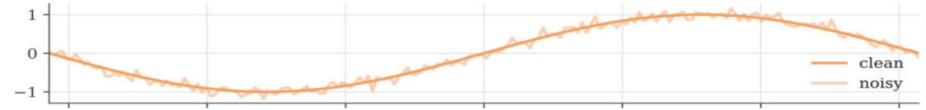
Image

Layer derivative



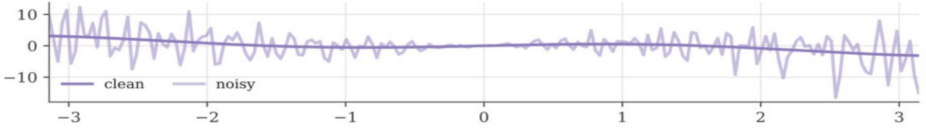
=

Effective transform



+

Operator variation

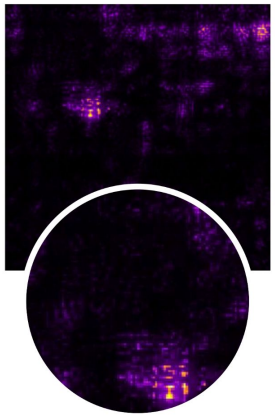


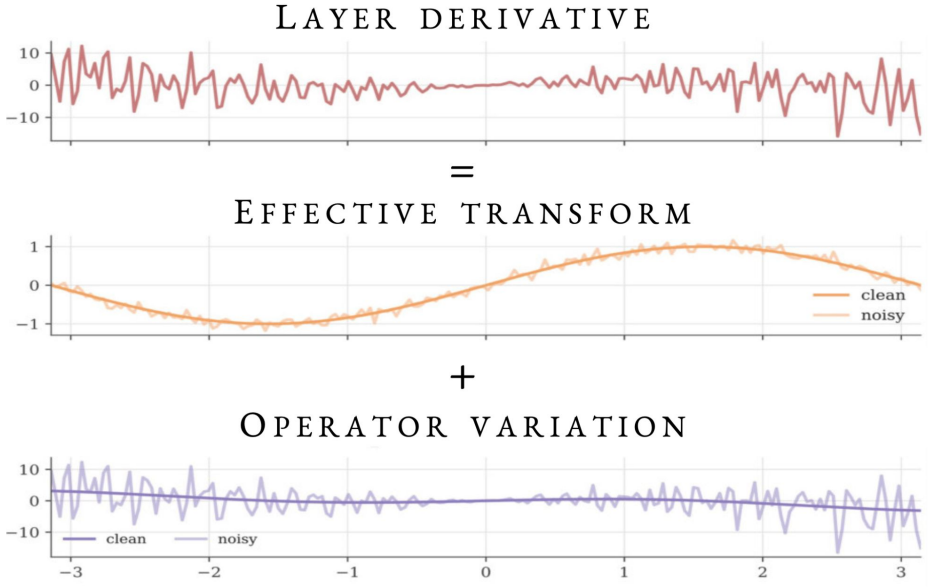
DAVE Construction

Step I: Effective Transformation

$$D_{\boldsymbol{X}}F = L(\boldsymbol{X}) + \big((D_{\boldsymbol{X}}L(\boldsymbol{X})(\cdot))\,\boldsymbol{X}\big)$$

layer derivative | effective transformation | operator variation

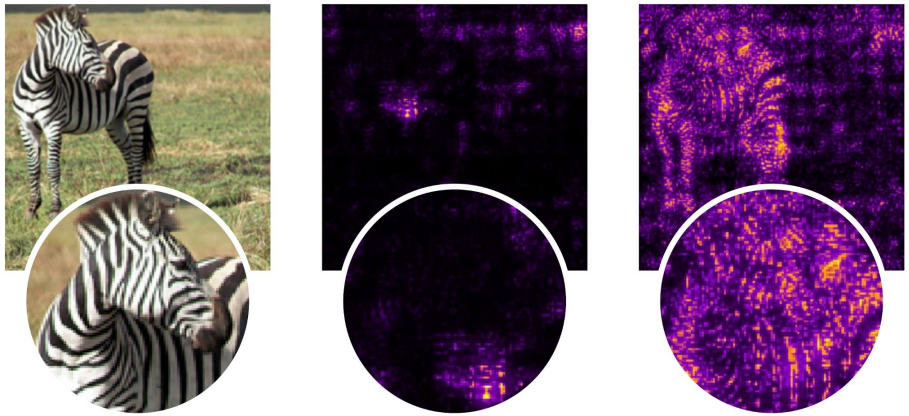


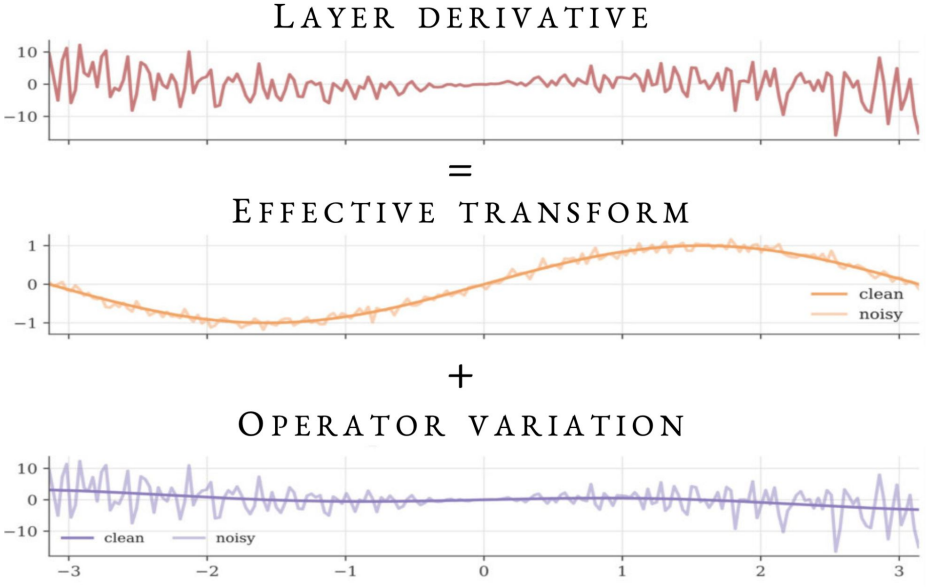


DAVE Construction

Step I: Effective Transformation

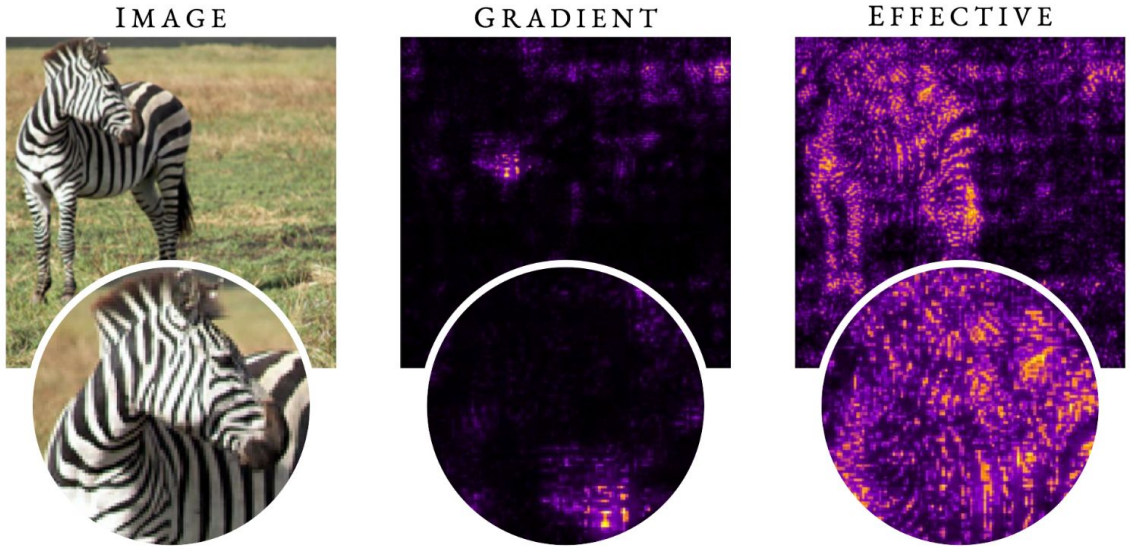
$$\underbrace{D_{\boldsymbol{X}}F}_{\text{layer derivative}} = \underbrace{L(\boldsymbol{X})}_{\text{effective transformation}} + \underbrace{((D_{\boldsymbol{X}}L(\boldsymbol{X})(\cdot))\,\boldsymbol{X})}_{\text{operator variation}}$$





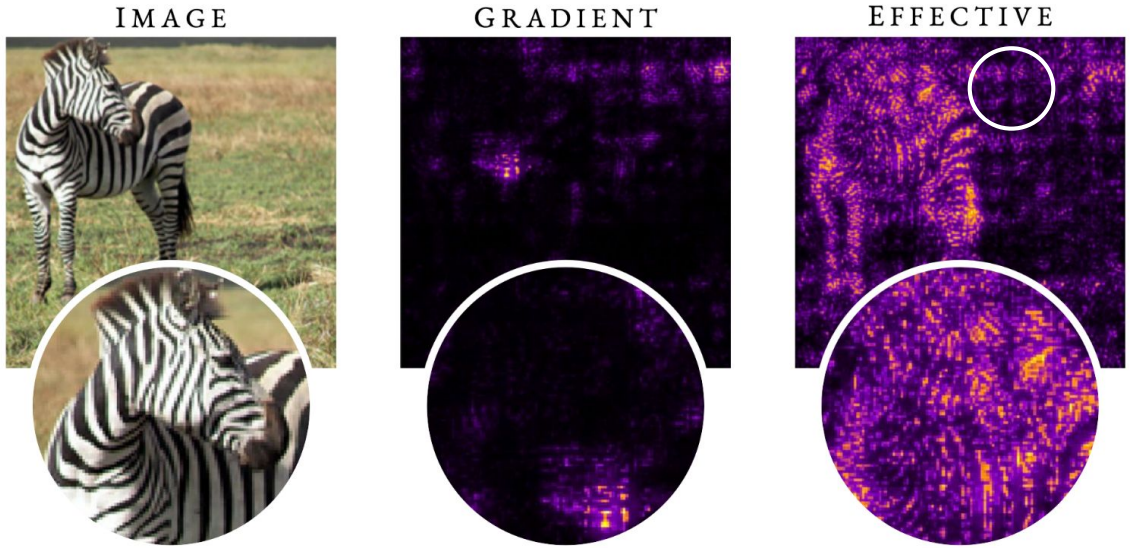
DAVE Construction

Step II: Equivariant Transformation



DAVE Construction

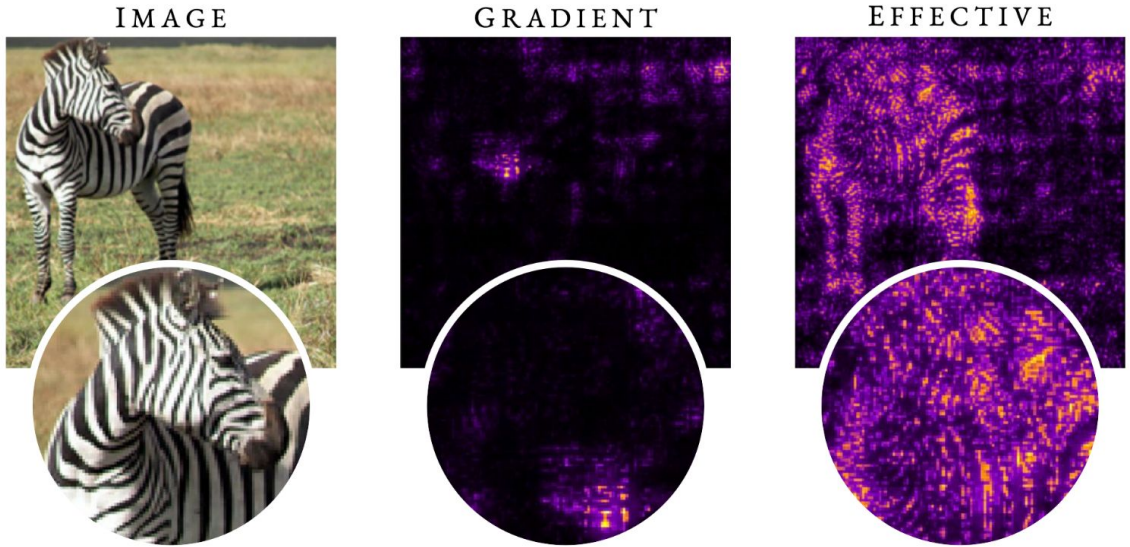
Step II: Equivariant Transformation



DAVE Construction

Step II: Equivariant Transformation

$$\boldsymbol{W}_L^{eq}(\boldsymbol{X}) = \int_G [\tau^{-1} \circ \boldsymbol{W}_L \circ \tau](\boldsymbol{X})\, d\nu(\tau)$$

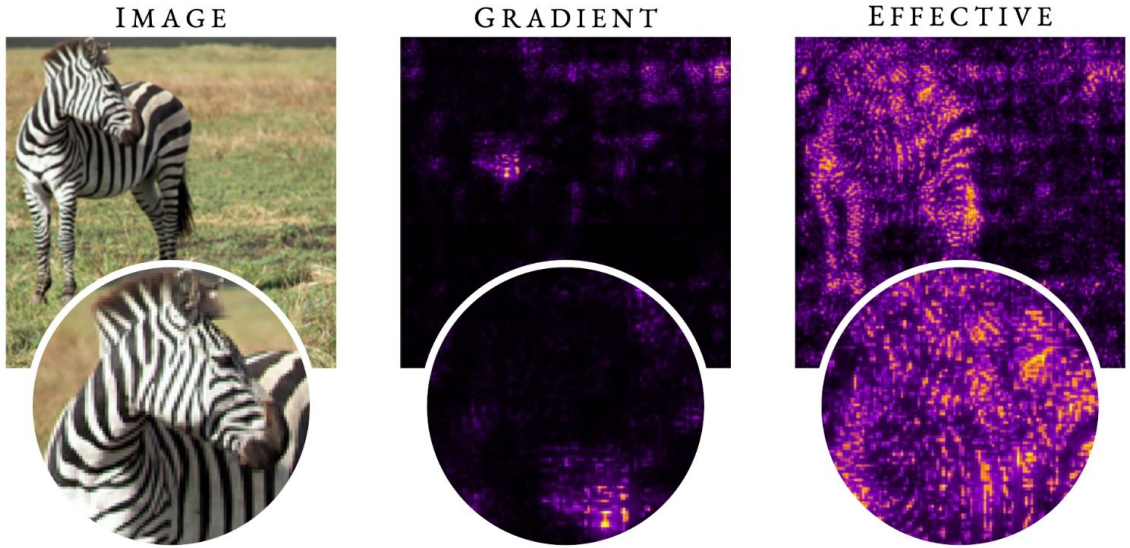


DAVE Construction

Step II: Equivariant Transformation

$$\boldsymbol{W}_L^{eq}(\boldsymbol{X}) = \int_G [\tau^{-1} \circ \boldsymbol{W}_L \circ \tau](\boldsymbol{X})\, d\nu(\tau)$$

Rotations / Translations

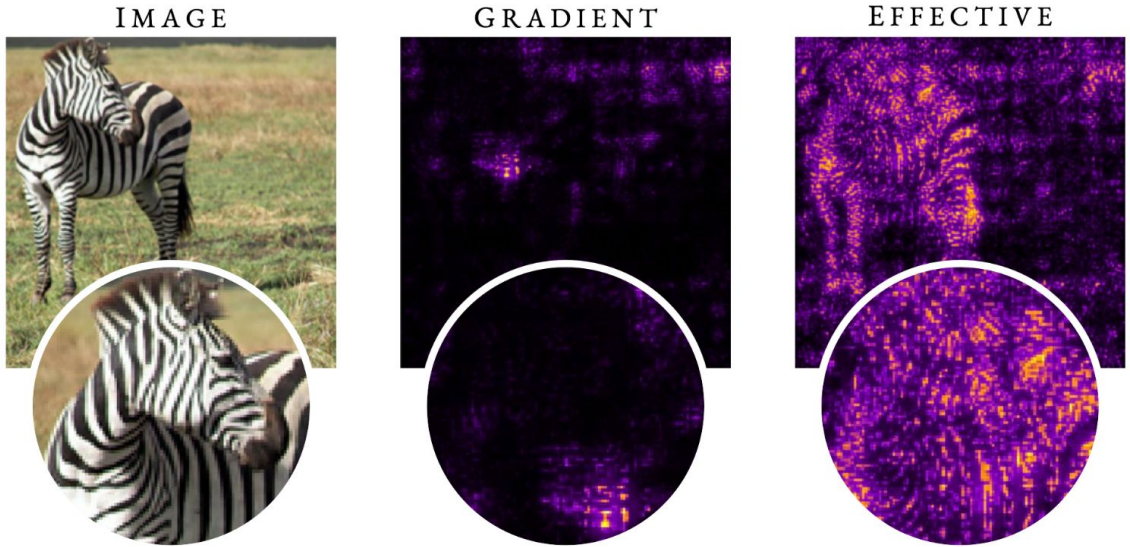


DAVE Construction

Step II: Equivariant Transformation

$$\boldsymbol{W}_L^{eq}(\boldsymbol{X}) = \int_G [\tau^{-1} \circ \boldsymbol{W}_L \circ \tau](\boldsymbol{X})\, d\nu(\tau)$$

Rotations / Translations Local neighborhood

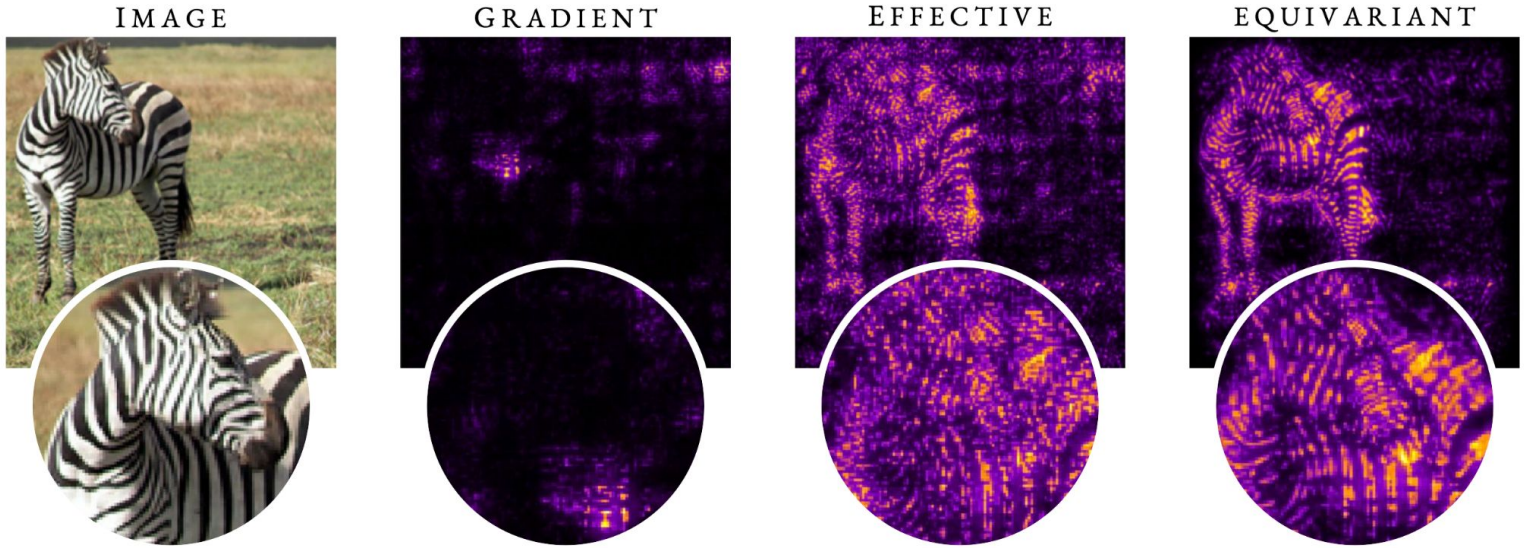


DAVE Construction

Step II: Equivariant Transformation

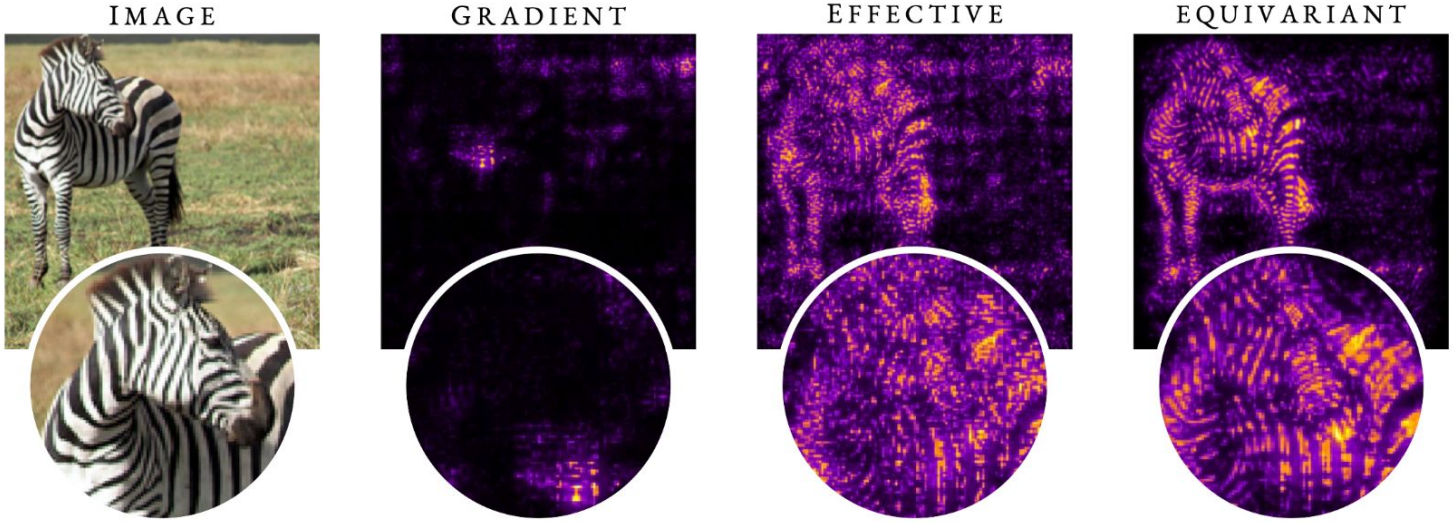
$$\boldsymbol{W}_L^{eq}(\boldsymbol{X}) = \int_G [\tau^{-1} \circ \boldsymbol{W}_L \circ \tau](\boldsymbol{X})\, d\nu(\tau)$$

Rotations / Translations Local neighborhood



DAVE Construction

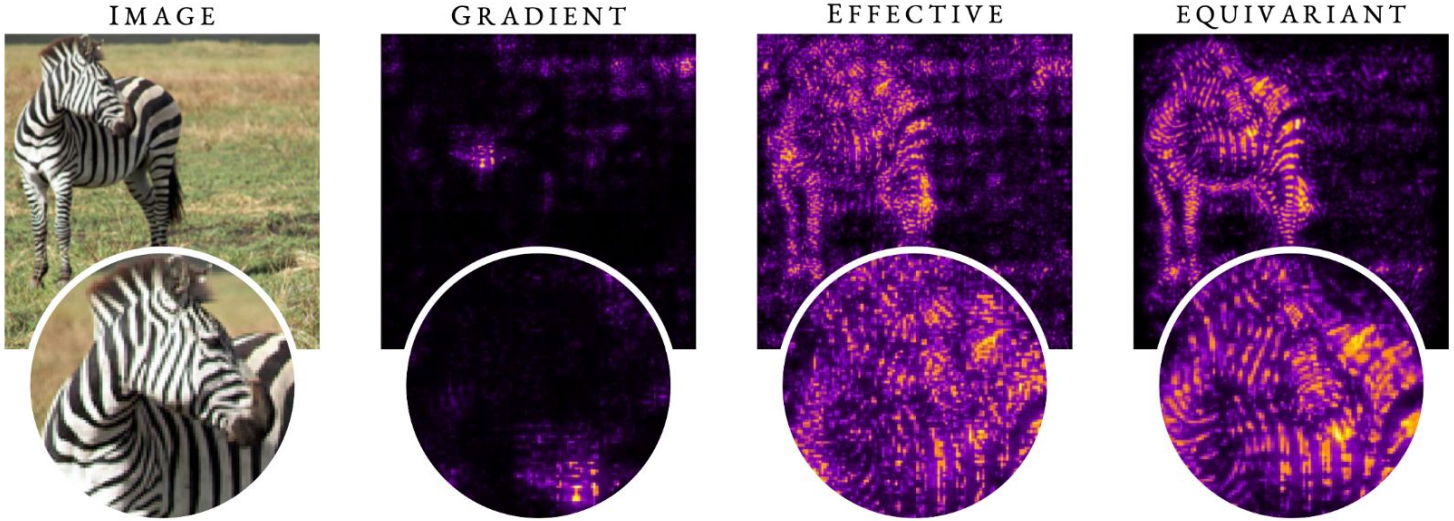
Step III: Low-pass Filter



DAVE Construction

Step III: Low-pass Filter

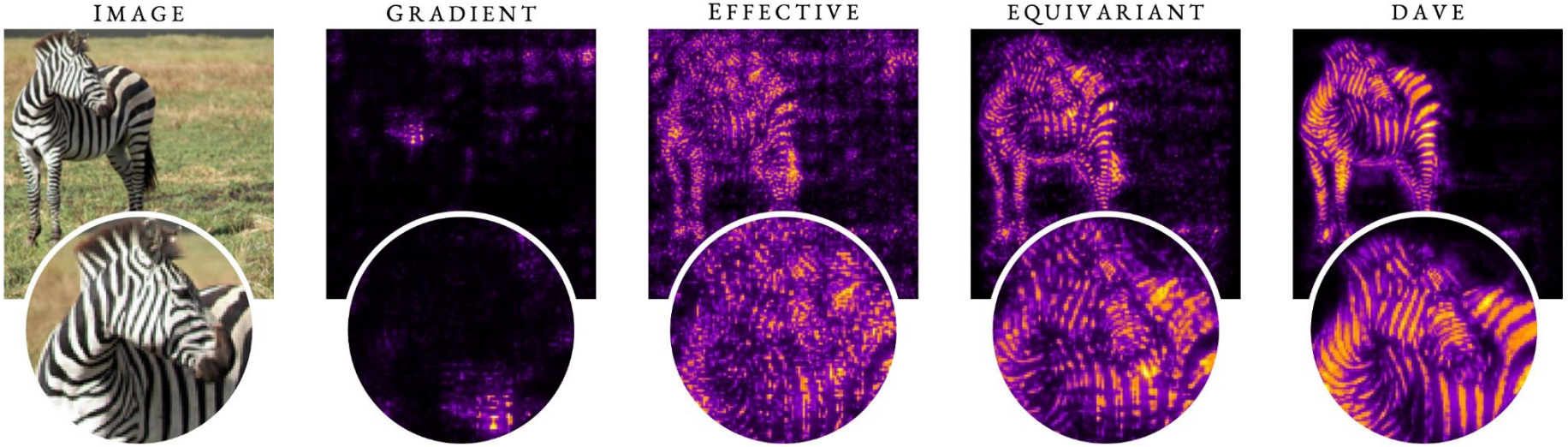
$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\Sigma)}\left[\boldsymbol{W}_L^{\mathrm{eq}}(\boldsymbol{X}+\boldsymbol{\epsilon})\right]$$



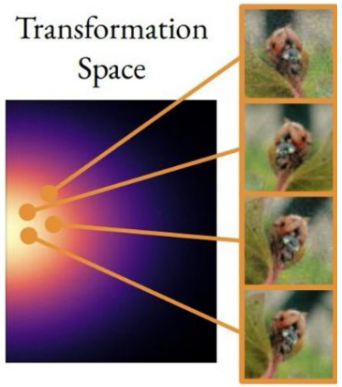
DAVE Construction

Step III: Low-pass Filter

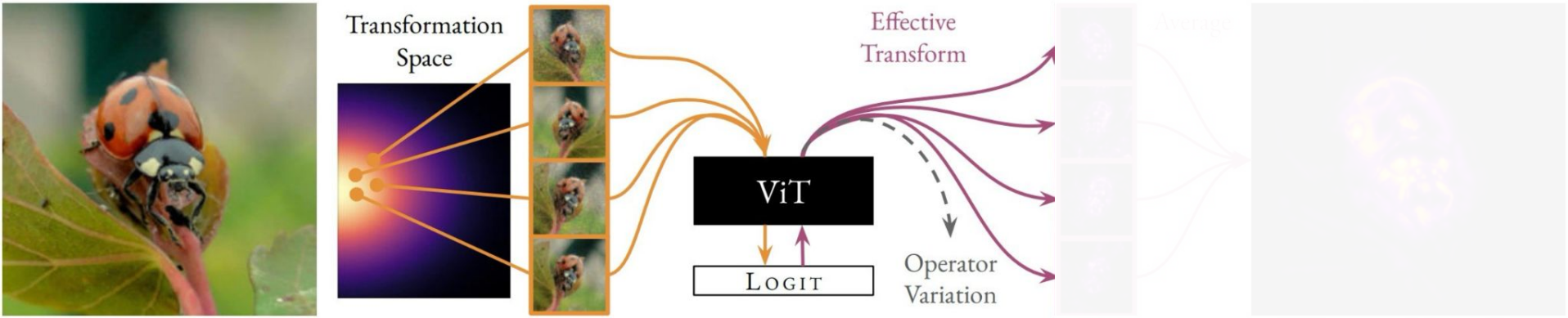
$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\Sigma)}\left[\boldsymbol{W}_L^{\text{eq}}(\boldsymbol{X}+\boldsymbol{\epsilon})\right]$$



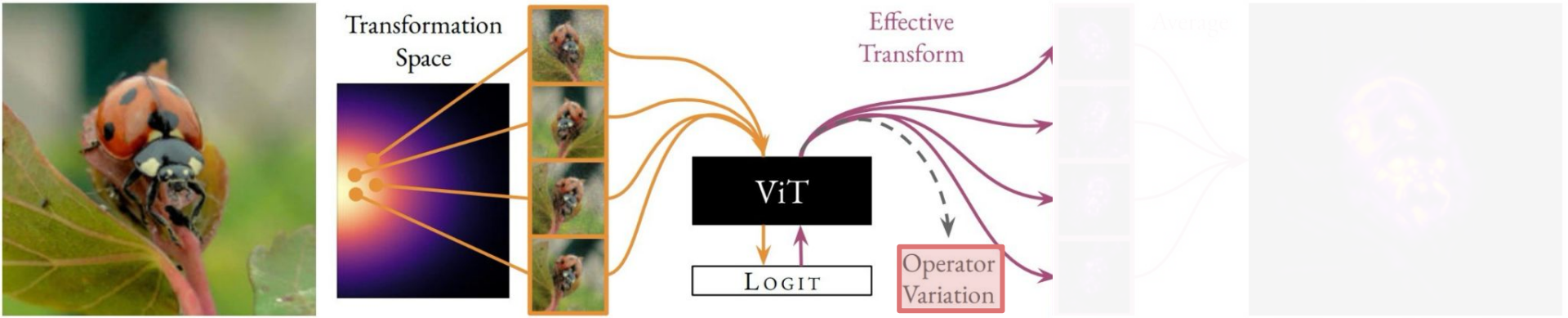
DAVE Implementation



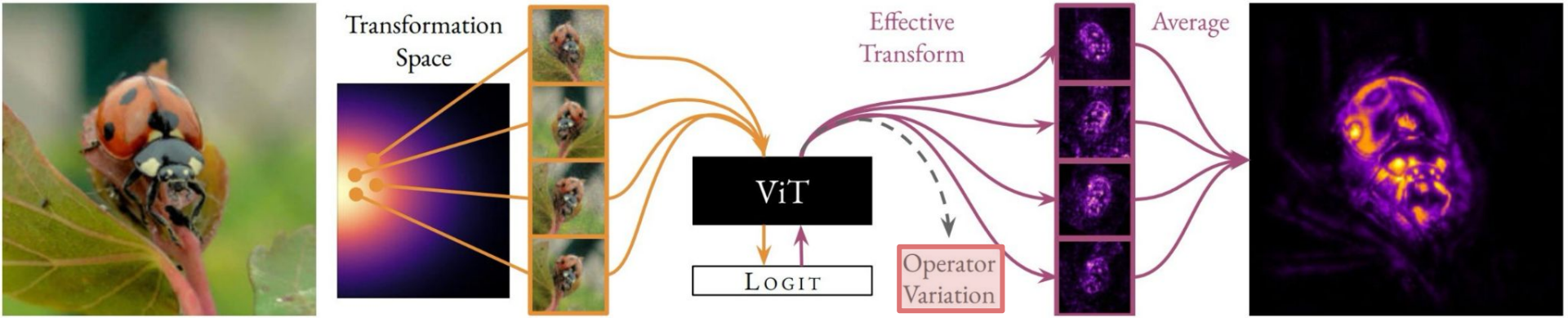
DAVE Implementation



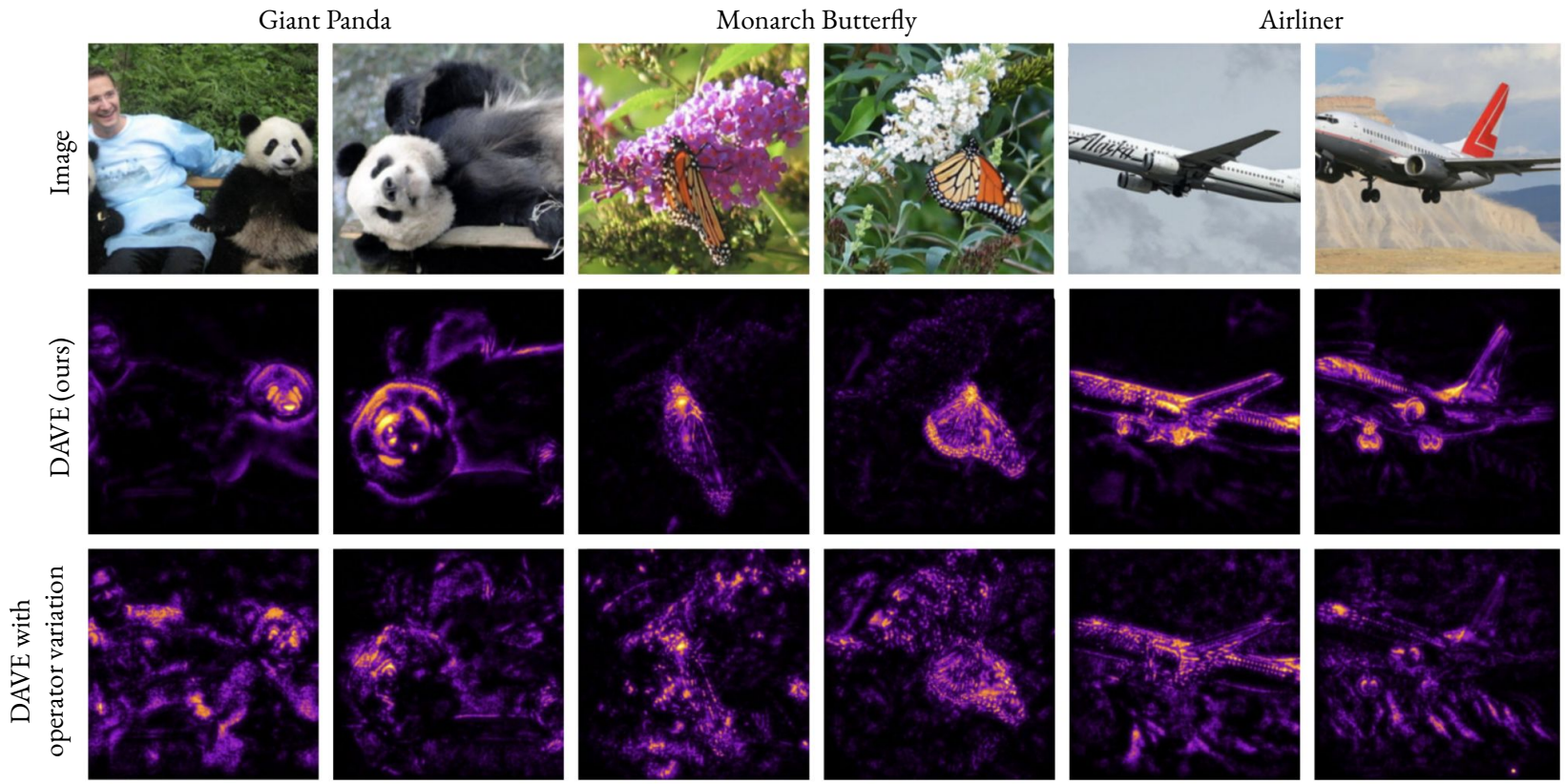
DAVE Implementation



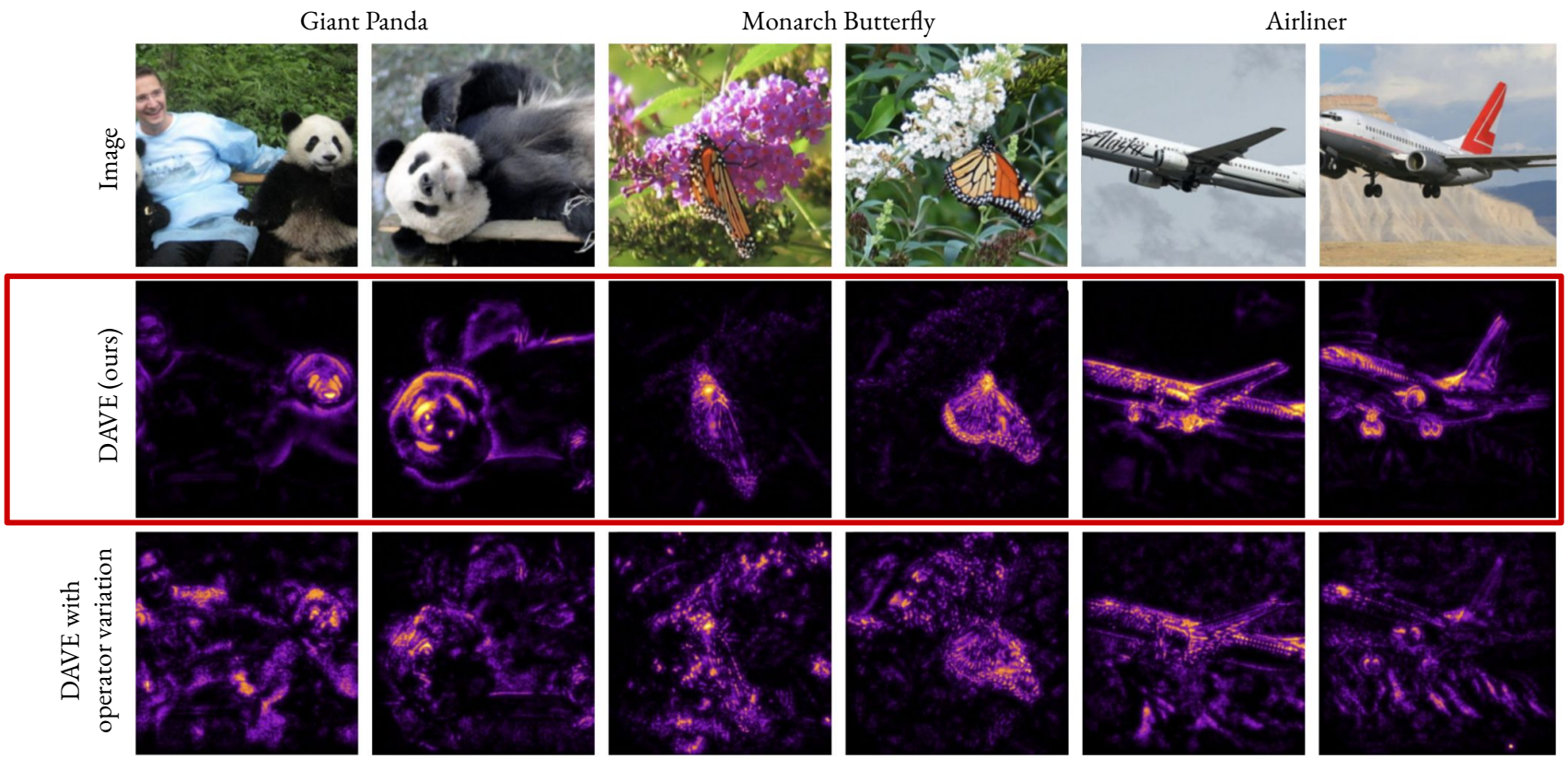
DAVE Implementation



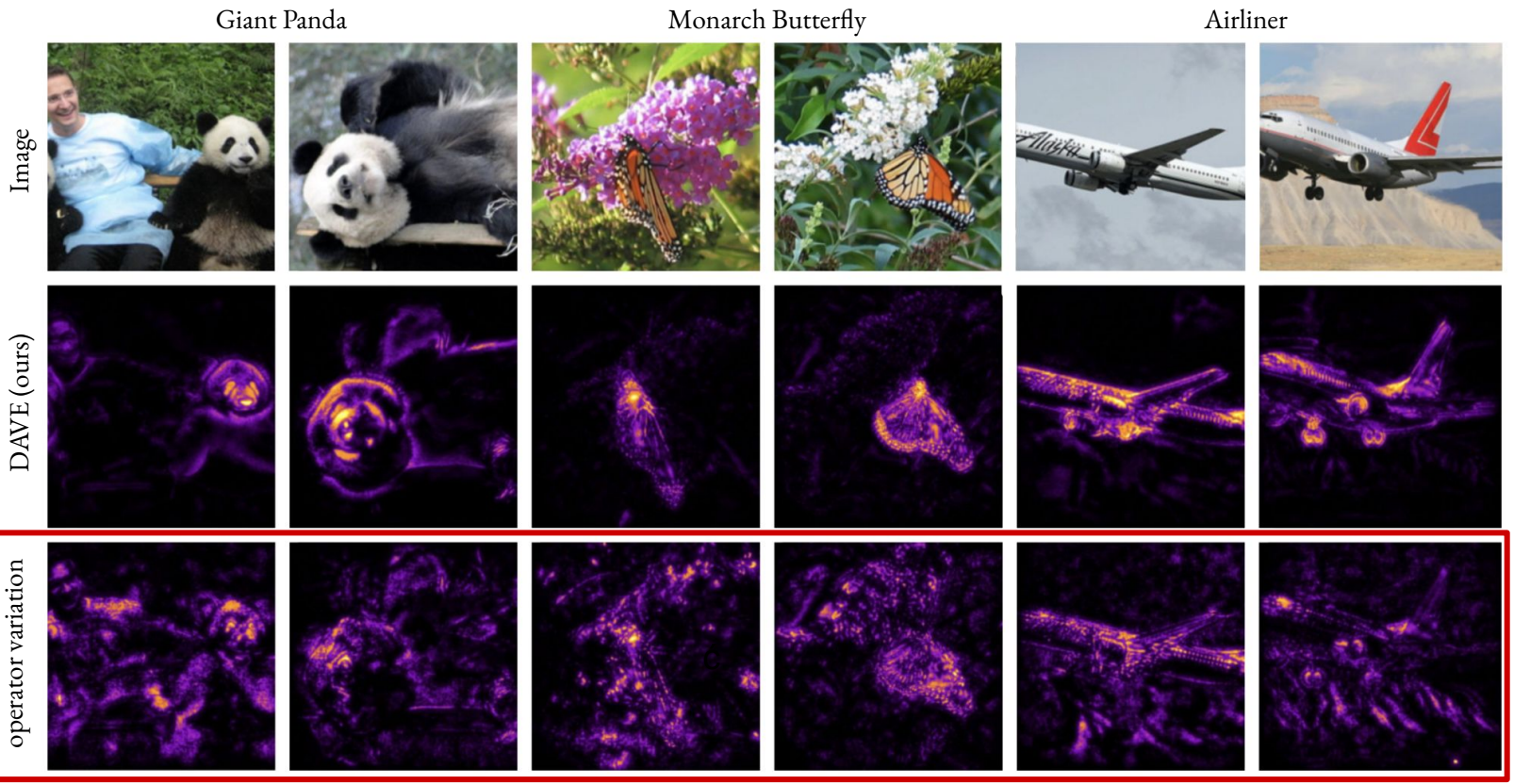
Operator Variation Effect



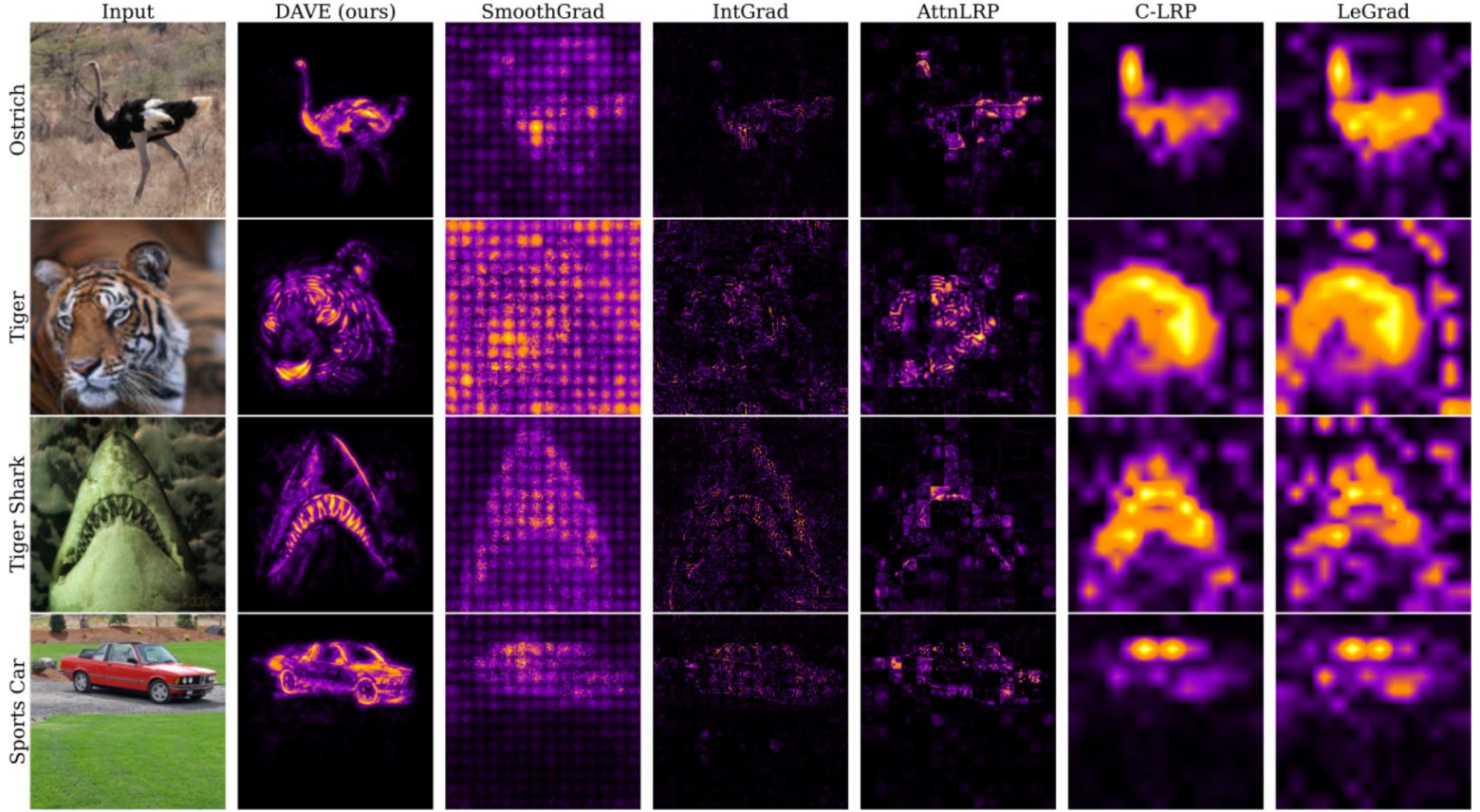
Operator Variation Effect



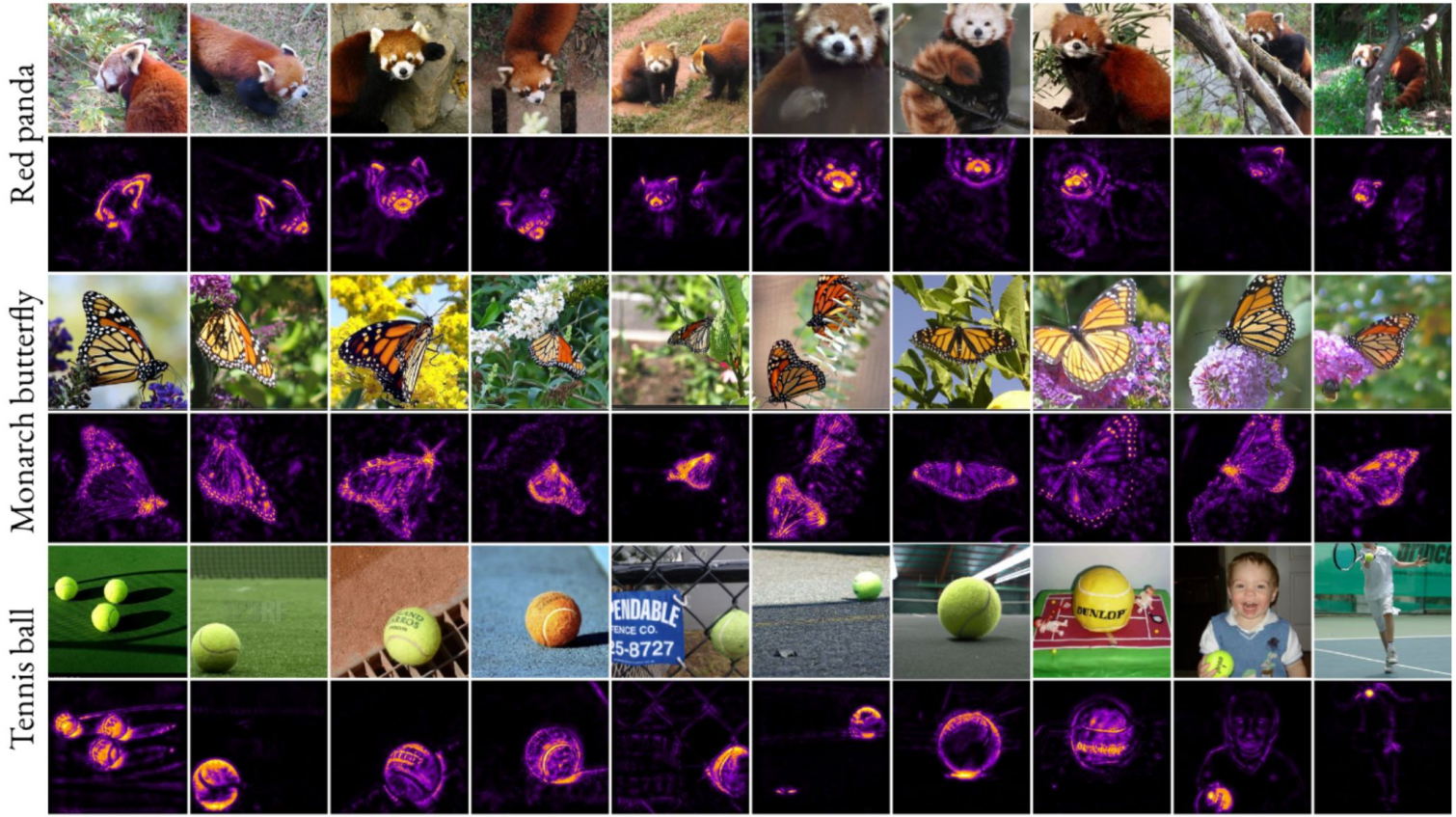
Operator Variation Effect



Qualitative Results

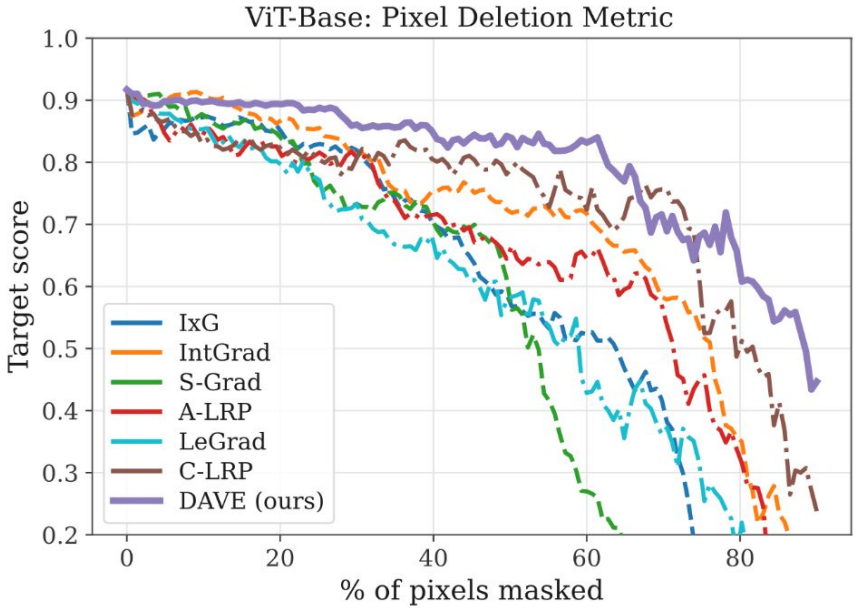


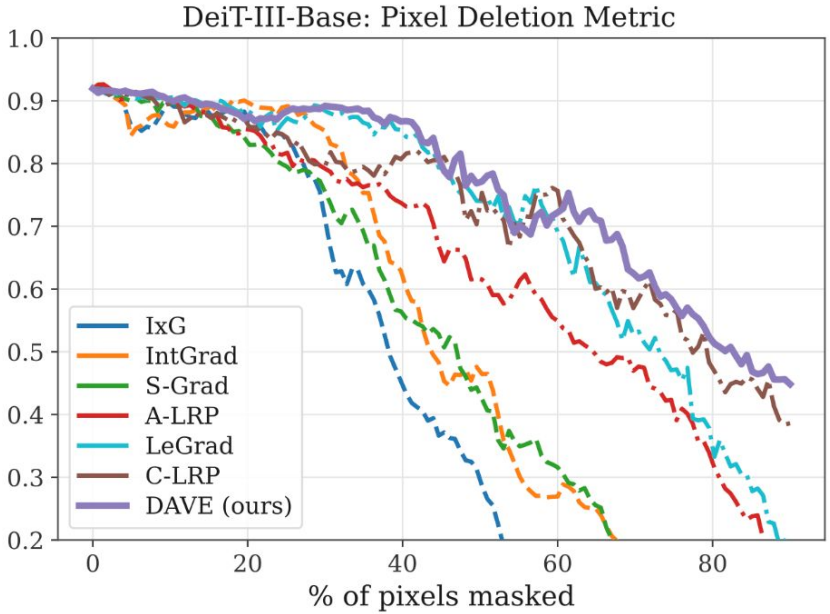
Class Consistency



Quantitative Results

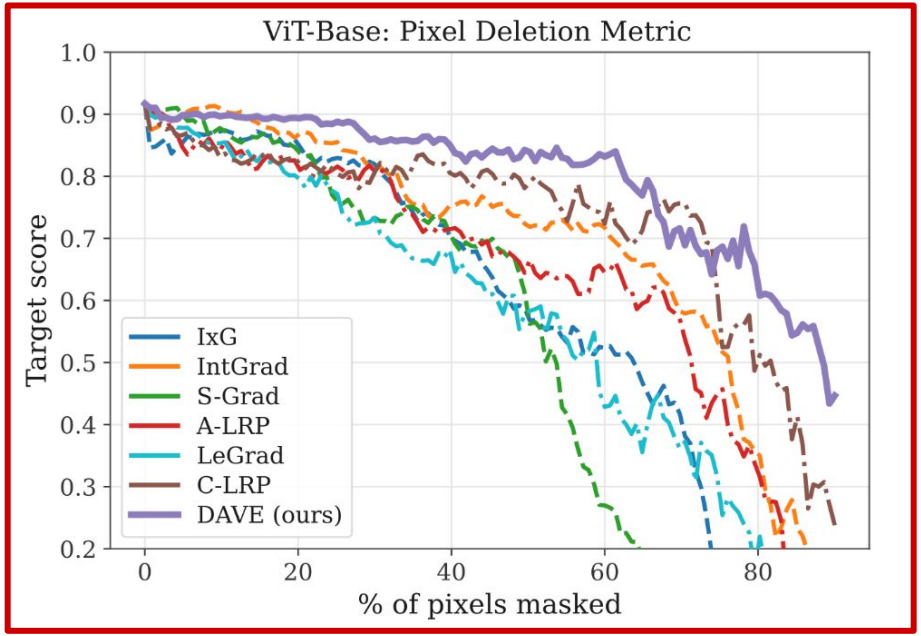
Pixel Deletion

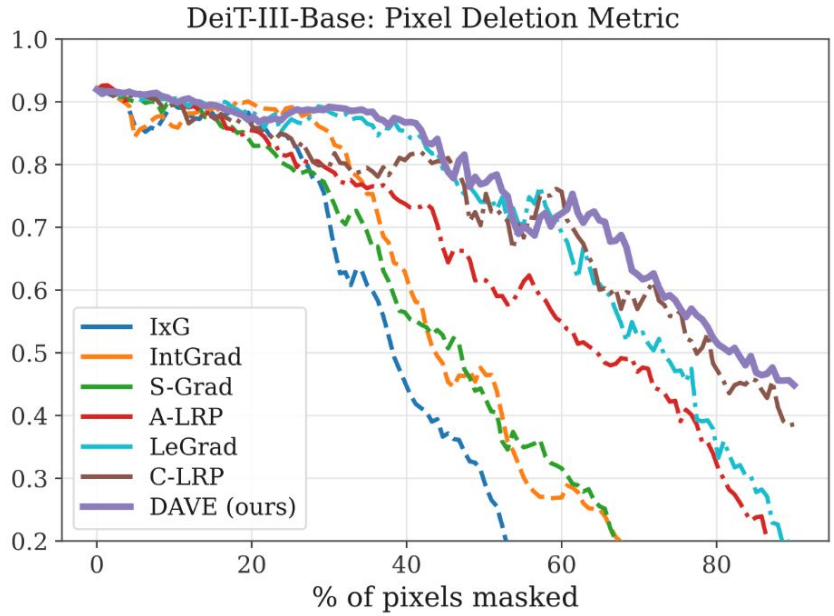




Quantitative Results

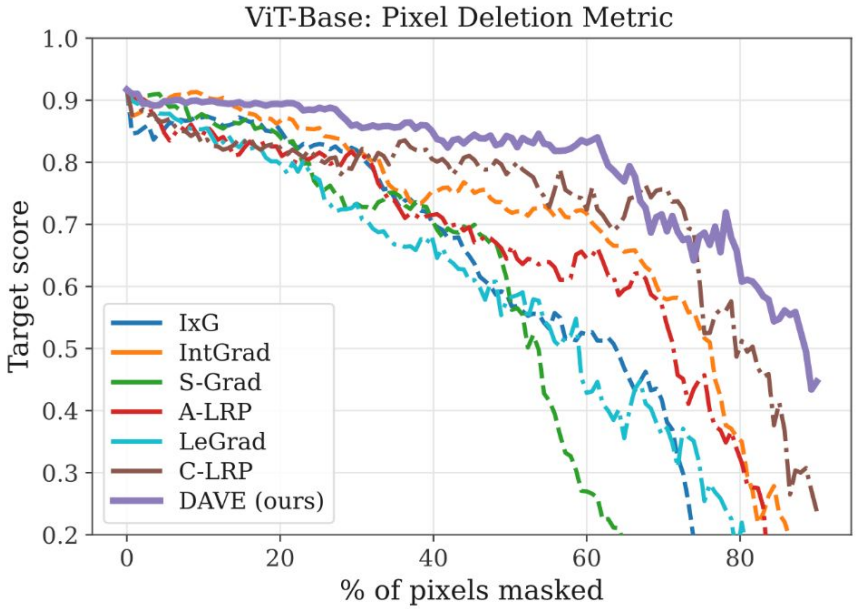
Pixel Deletion

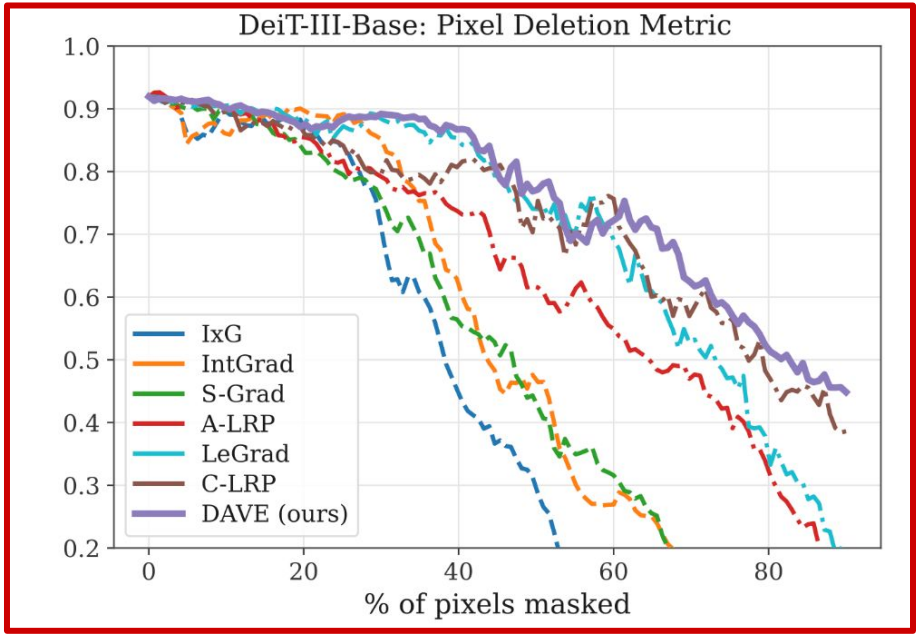




Quantitative Results

Pixel Deletion

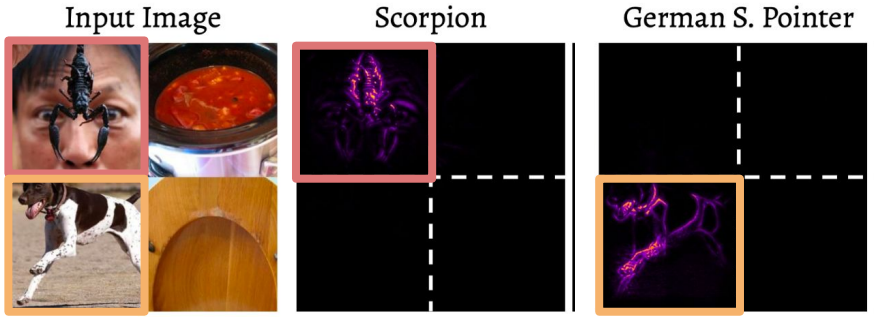




Quantitative Results

Localization

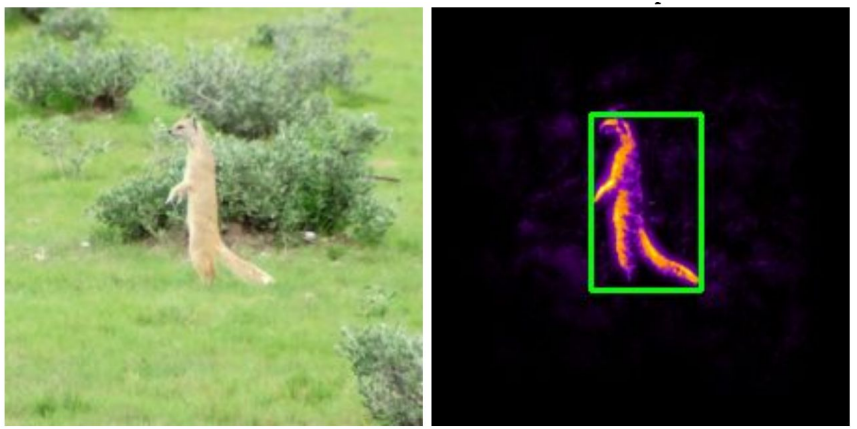
Method	GridPG (%) ViT-B	DeiT-B	D-III-B	DINO-B
I×G	32.67	30.25	30.01	33.28
IntGrad	39.86	36.11	31.68	36.98
S-Grad	34.27	30.18	31.48	33.13
LeGrad	47.71	42.58	34.62	28.96
A-LRP	58.40	54.63	53.84	37.49
C-LRP	54.98	55.47	52.27	49.99
DAVE (ours)	**60.19**	**63.52**	**65.76**	**51.33**
Δ	**+1.79**	**+8.05**	**+11.92**	**+1.35**



Quantitative Results

Localization

Method	EnergyPG (%) ViT-B	DeiT-B	D-III-B	DINO-B
I×G	55.33	62.72	67.32	69.98
IntGrad	58.51	64.22	68.12	72.15
S-Grad	56.02	60.78	68.10	70.93
LeGrad	80.06	77.83	77.54	82.26
A-LRP	60.75	68.16	77.65	75.98
C-LRP	**80.82**	79.62	81.94	81.56
DAVE (ours)	78.60	**82.23**	**82.43**	**83.38**
Δ	-2.22	+2.61	+0.49	+1.12



Takeaways

- ViT attributions suffer from **architecture-induced artifacts**.
- ViT gradients mix **effective transform** and **operator variation**.
- DAVE isolates **stable**, **transformation-consistent** signal from **effective transform**.
- DAVE achieves **state-of-the-art** pixel-level attribution quality.

Takeaways

- ViT attributions suffer from **architecture-induced artifacts**.
- ViT gradients mix **effective transform** and **operator variation**.
- DAVE isolates **stable, transformation-consistent** signal from **effective transform**.
- DAVE achieves **state-of-the-art** pixel-level attribution quality.

Takeaways

- ViT attributions suffer from **architecture-induced artifacts**.
- ViT gradients mix **effective transform** and **operator variation**.
- DAVE isolates **stable**, **transformation-consistent** signal from **effective transform**.
- DAVE achieves **state-of-the-art** pixel-level attribution quality.

Takeaways

- ViT attributions suffer from **architecture-induced artifacts**.
- ViT gradients mix **effective transform** and **operator variation**.
- DAVE isolates **stable**, **transformation-consistent** signal from **effective transform**.
- DAVE achieves **state-of-the-art** pixel-level ViT attribution quality.