Coherentice: Invertible Concept-based Explainability Framework For CNNs Beyond Fidelity

Ugochukwu Ejike Akpudo, Yongsheng Gao, Jun Zhou, Andrew Lewis

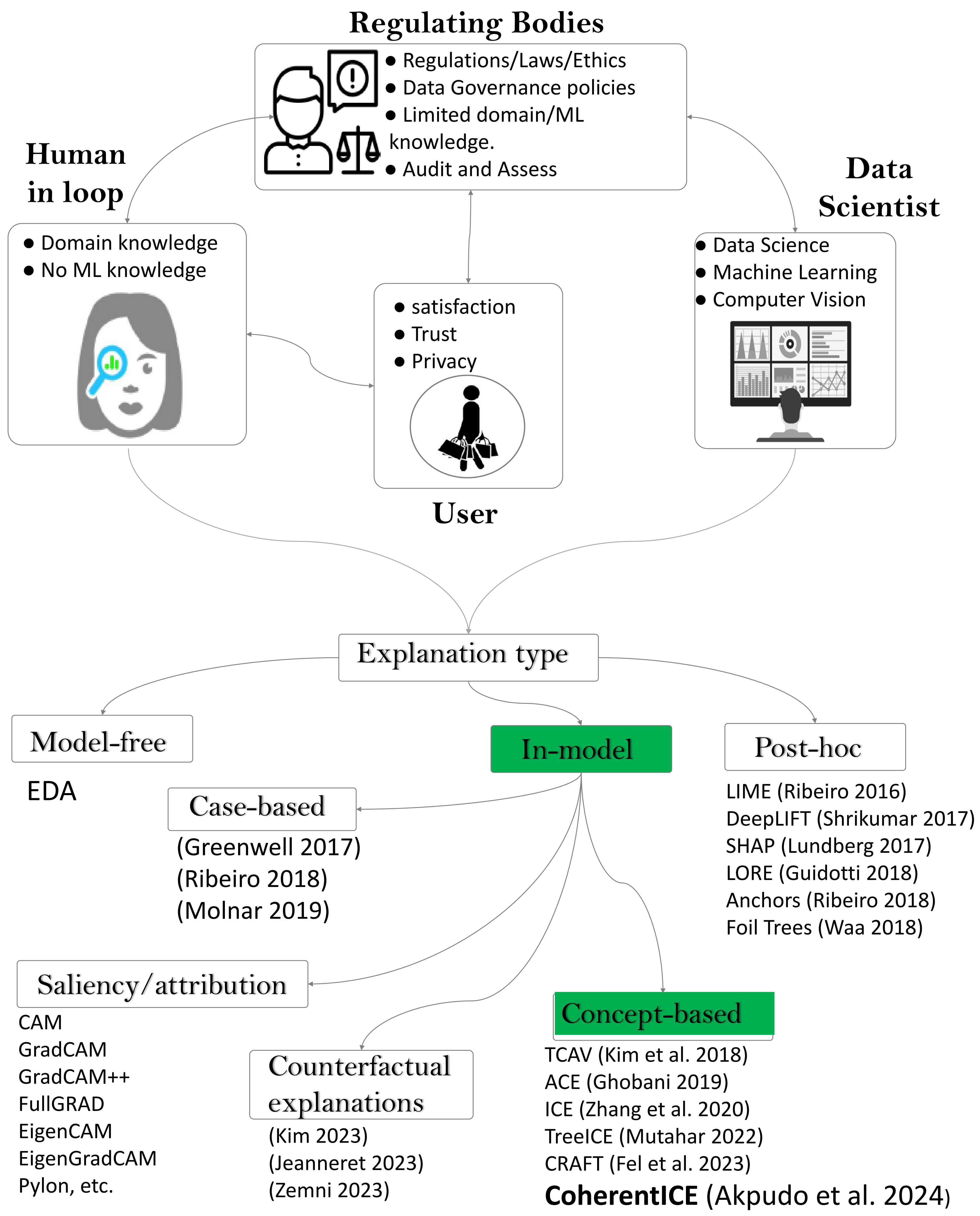
Griffith UNIVERSITY
Queensland, Australia

Motivation

> "While existing methods reveal **what a CNN saw** (as concepts & prototypes), it is imperative to evaluate not only how accurate the concepts are but also, **how consistent the explanations are**."

Objective

> "Investigate the meaningfulness of concept explanations using a novel faithfulness evaluation paradigm: the Faith score, for accuracy and consistency."



Regulating Bodies
- Regulations/Laws/Ethics
- Data Governance policies
- Limited domain/ML knowledge.
- Audit and Assess

Human in loop
- Domain knowledge
- No ML knowledge

Data Scientist
- Data Science
- Machine Learning
- Computer Vision

- satisfaction
- Trust
- Privacy

User

Explanation type

Model-free

EDA

In-model

Post-hoc

LIME (Ribeiro 2016)
DeepLIFT (Shrikumar 2017)
SHAP (Lundberg 2017)
LORE (Guidotti 2018)
Anchors (Ribeiro 2018)
Foil Trees (Waa 2018)

Case-based

(Greenwell 2017)
(Ribeiro 2018)
(Molnar 2019)

Saliency/attribution

CAM
GradCAM
GradCAM++
FullGRAD
EigenCAM
EigenGradCAM
Pylon, etc.

Counterfactual explanations

(Kim 2023)
(Jeanneret 2023)
(Zemni 2023)

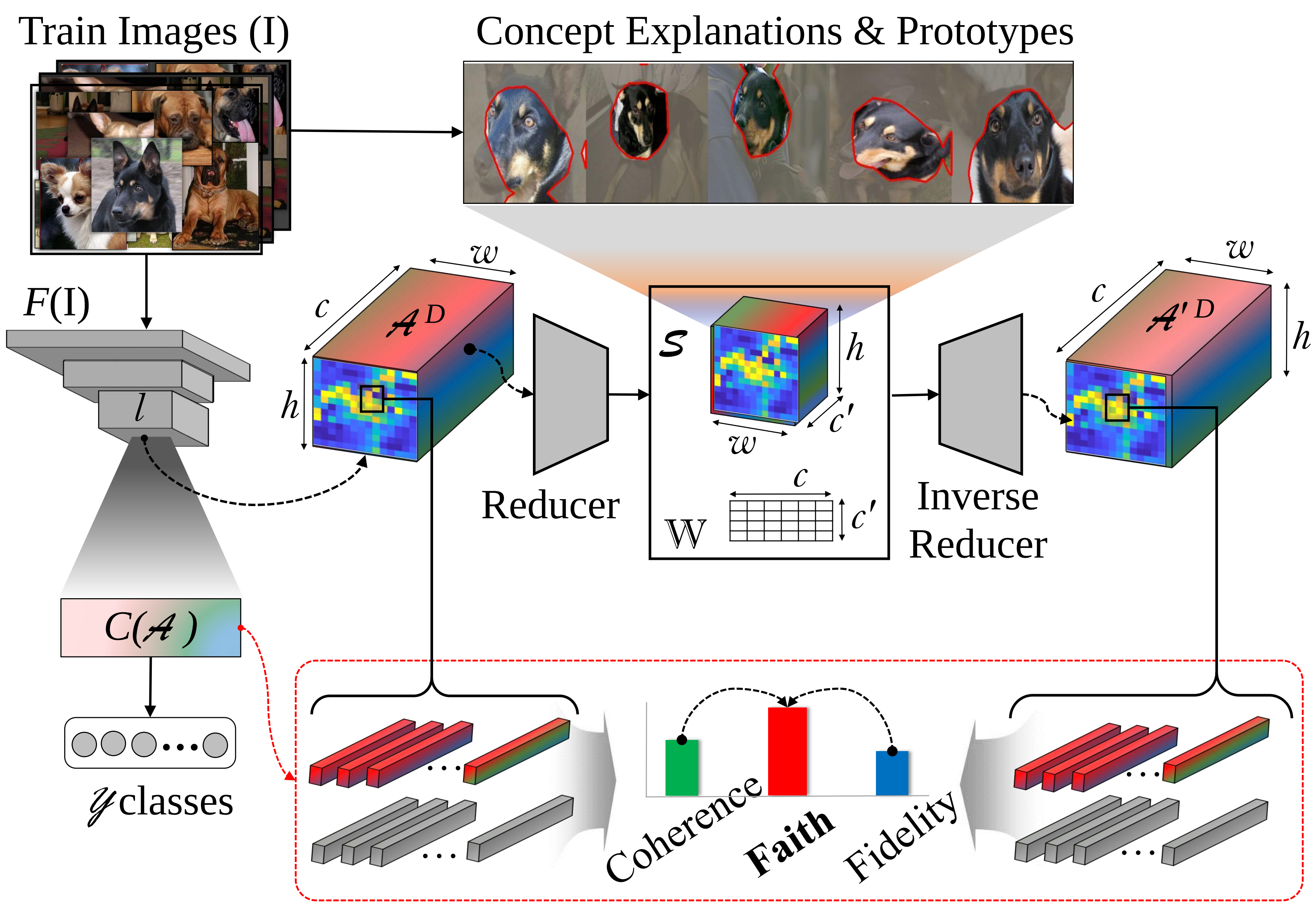
Concept-based

TCAV (Kim et al. 2018)
ACE (Ghobani 2019)
ICE (Zhang et al. 2020)
TreeICE (Mutahar 2022)
CRAFT (Fel et al. 2023)
CoherentICE (Akpudo et al. 2024)

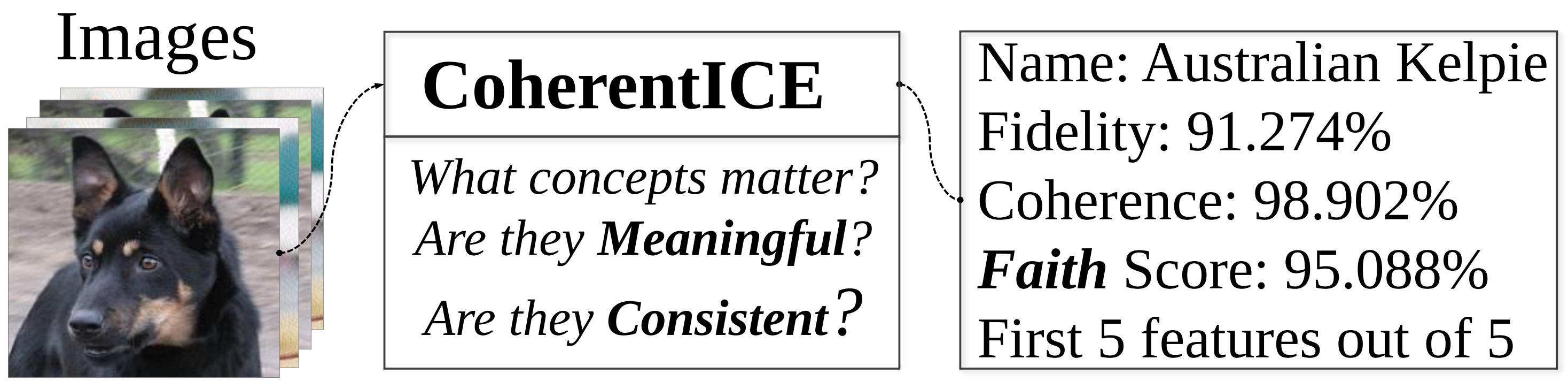
Pseudocode for CoherentICE

```
Input: Image ($I$), CNN backend ($F(I)$), NMF reducer ($\mathbb{N}$)
Parameter: User-defined no. of concepts ($c'$), threshold ($\lambda$)
Output: Faith score ($^*\mathbb{F}_{C\leftrightarrow R}$), Weight ($P\omega_{C,m,l}$), and FeatureRank
Split $\{F(I)|F(I) = E(I) \cdot C(A_l^D)\}$ such that
$\mathcal{A}_l^D \in \mathbb{R}^{n\times b\times w\times c}$ in layer $l$;
Flatten $\mathcal{A}_l^D$ to $\mathcal{G} \in \mathbb{R}^{(n\times b\times w)\times c}$;
for all $y_i$ in $Y$ classes do
    for all $g^{(i,j)}$ at $l$ in $\mathcal{G}$ do
        Transform $a^{(i,j)}$ with $\mathbb{N}$ such that $V = SP + u$
        where $\{S \in \mathbb{R}^{(n\times b\times w)\times c'}, P \in \mathbb{R}^{c\times c'}\}$;
        Create heatmap $\{i \in I, s \in S \mid E(I) \equiv \hat{E}_\lambda(s)\}$
        Compute Weight using Proposition 2
    end for
    Collate prototypes $\hat{E}_{\lambda,y_i}(s)$;
    Sort Weight to produce FeatureRank;
    Invert $\mathbb{N}'(S^d, P) \longrightarrow \mathcal{A}_l'^D \in \mathbb{R}^{n\times b\times w\times c}$;
    Compute $^*\mathbb{F}_{C\leftrightarrow R}$ using Proposition 4.
end for
```

Proposed CoherentICE Framework



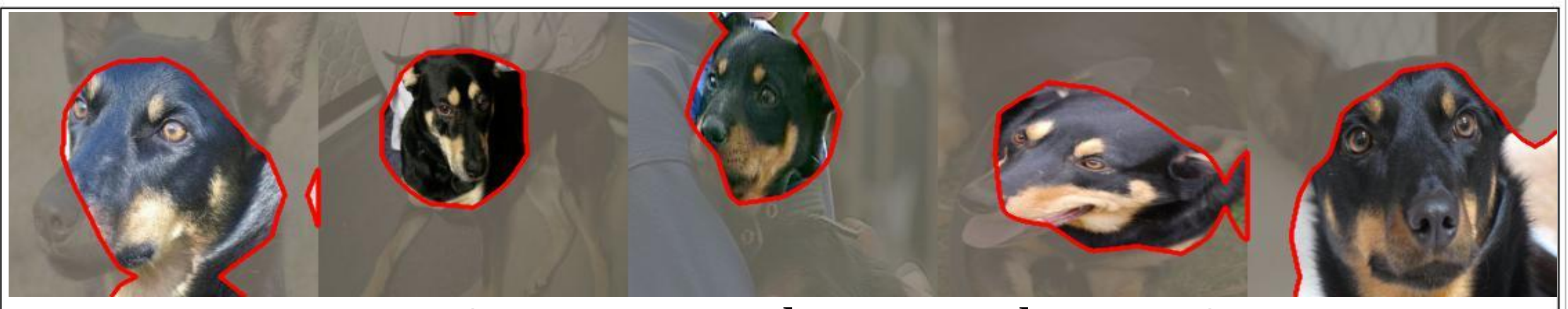
Application and Results



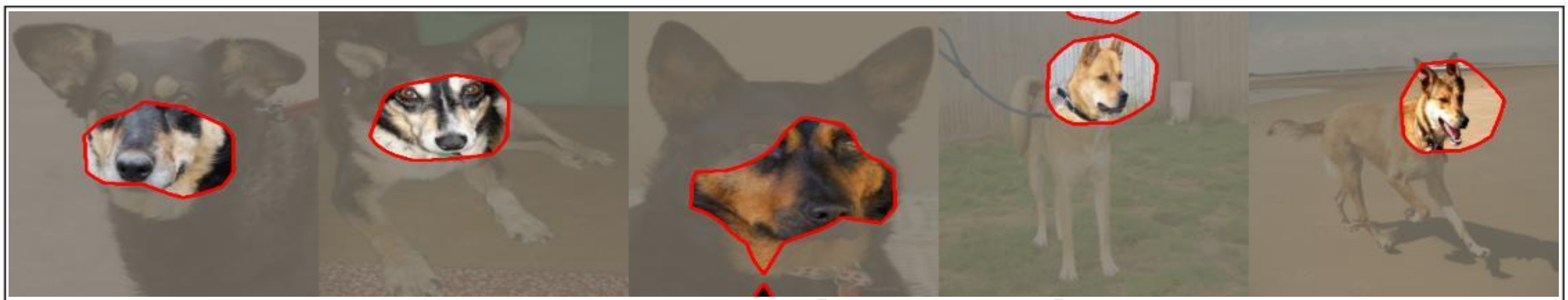
CoherentICE

What concepts matter?
*Are they **Meaningful**?*
*Are they **Consistent**?*

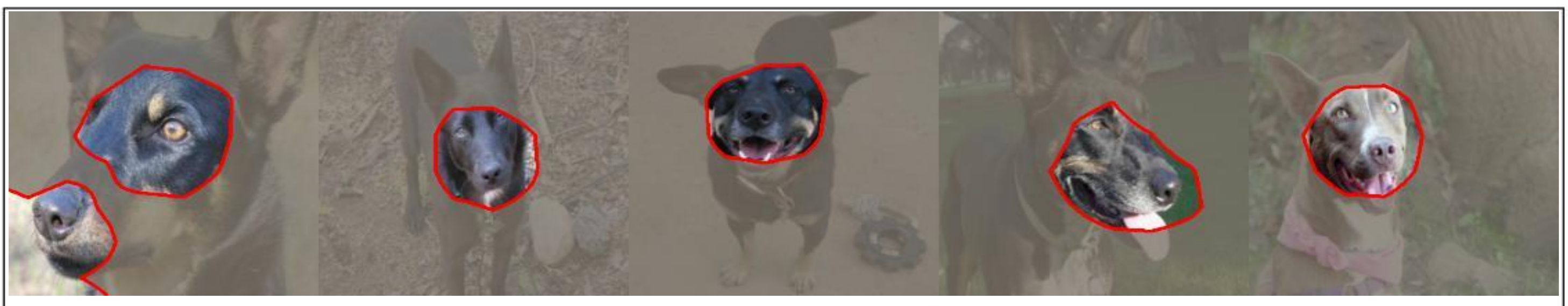
Name: Australian Kelpie
Fidelity: 91.274%
Coherence: 98.902%
Faith Score: 95.088%
First 5 features out of 5



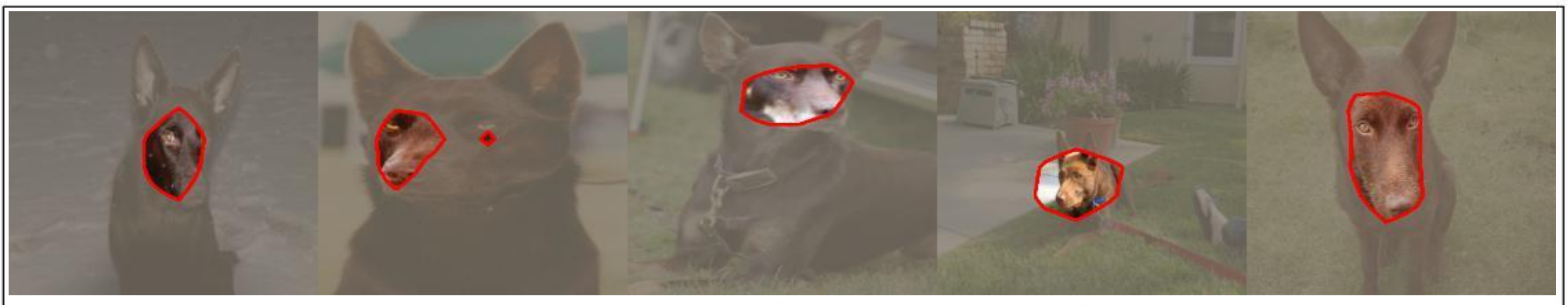
FeatureID: 0, FeatureRank: 1, Weight: 41.907
Concept: *Patterned grey hair on face*



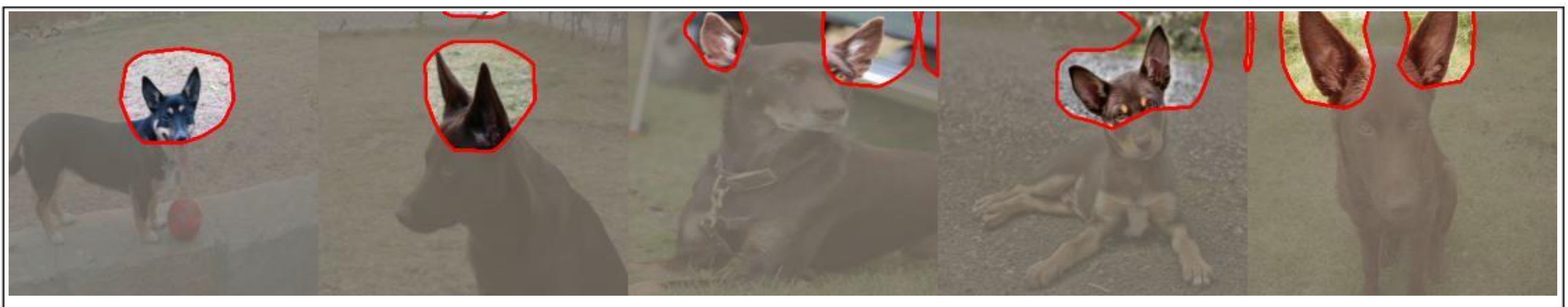
FeatureID: 2, FeatureRank: 2, Weight: 27.703
Concept: *Long narrow face with dark nose*



FeatureID: 1, FeatureRank: 3, Weight: 23.056
Concept: *Grey hair on eyebrow*



FeatureID: 3, FeatureRank: 4, Weight: 19.358
Concept: *Feisty eyes*



FeatureID: 4, FeatureRank: 5, Weight: 1.112
Concept: *Perky ears*

Significance of Study

> "*Imperfect concepts can be accepted as prototypes if they are consistent. However, trust in concepts diminishes if they are both imperfect and inconsistent.* ***Concepts can be trusted if they are consistently accurate.***"