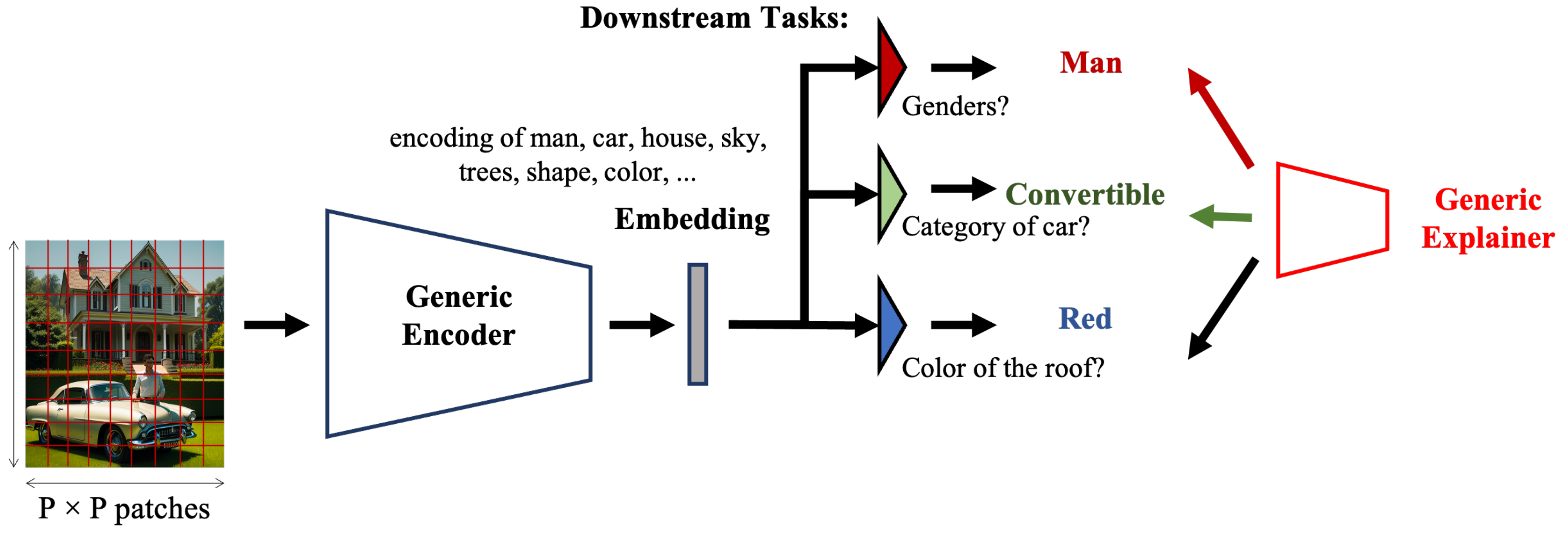
TVE: Learning Meta-attribution for Transferable Vision Explainer

Guanchu Wang[1], Yu-Neng Chuang[1], Fan Yang[2], Mengnan Du[3], Chia-Yuan Chang[4],Shaochen Zhong[1], Zirui Liu[1], Zhaozhuo Xu[5], Kaixiong Zhou[6], Xuanting Cai[7], Xia Hu[1]

[1] Rice University, [2] Wake Forest University, [3] New Jersey Institute of Technology, [4] Texas A&M University, [5] Stevens Institute of Technology, [6] North Carolina State University, [7] Meta Platforms, Inc.

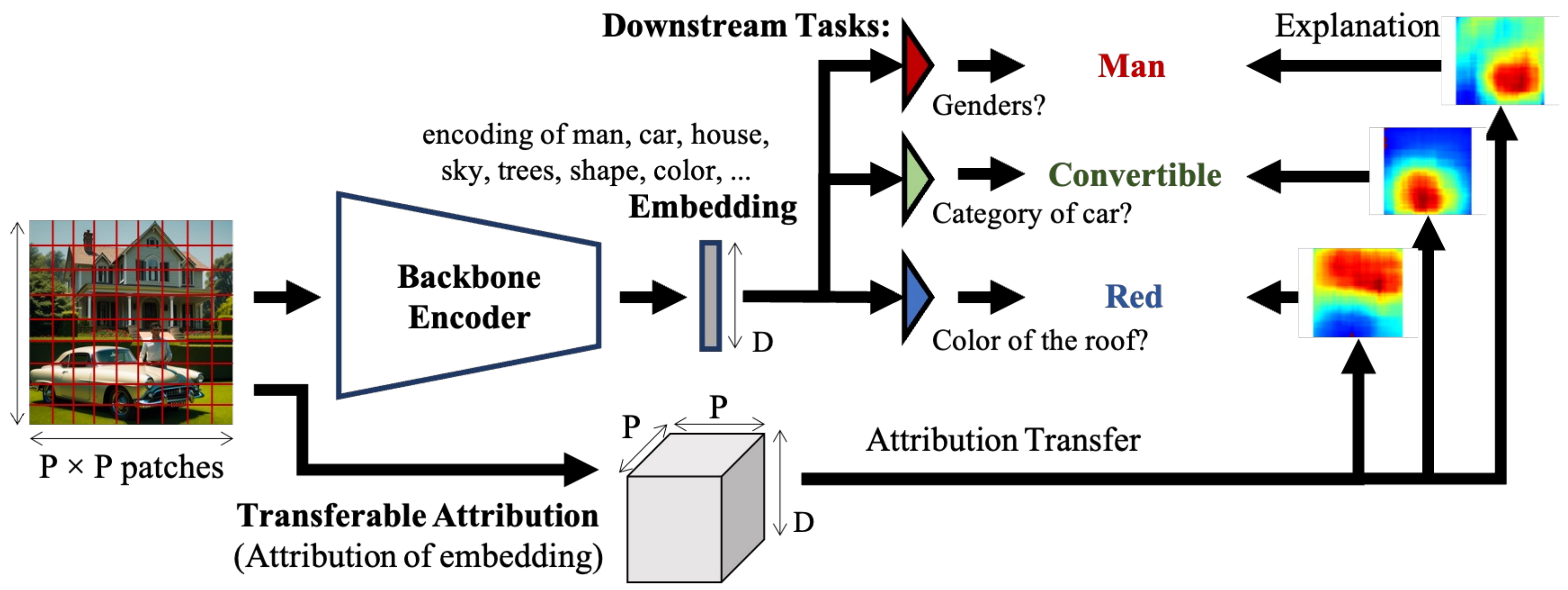
1.1 Transferable Vision Explainer (TVE)

Existing work is constrained to explaining the behavior of individual model predictions, and lacks the ability to transfer the explanation across various models and tasks. This limitation results in explaining various tasks being time- and resource-consuming. The primary goal of TVE is to achieve transferability through a pre-training process on large-scale image datasets, such that it can seamlessly explain various downstream tasks.



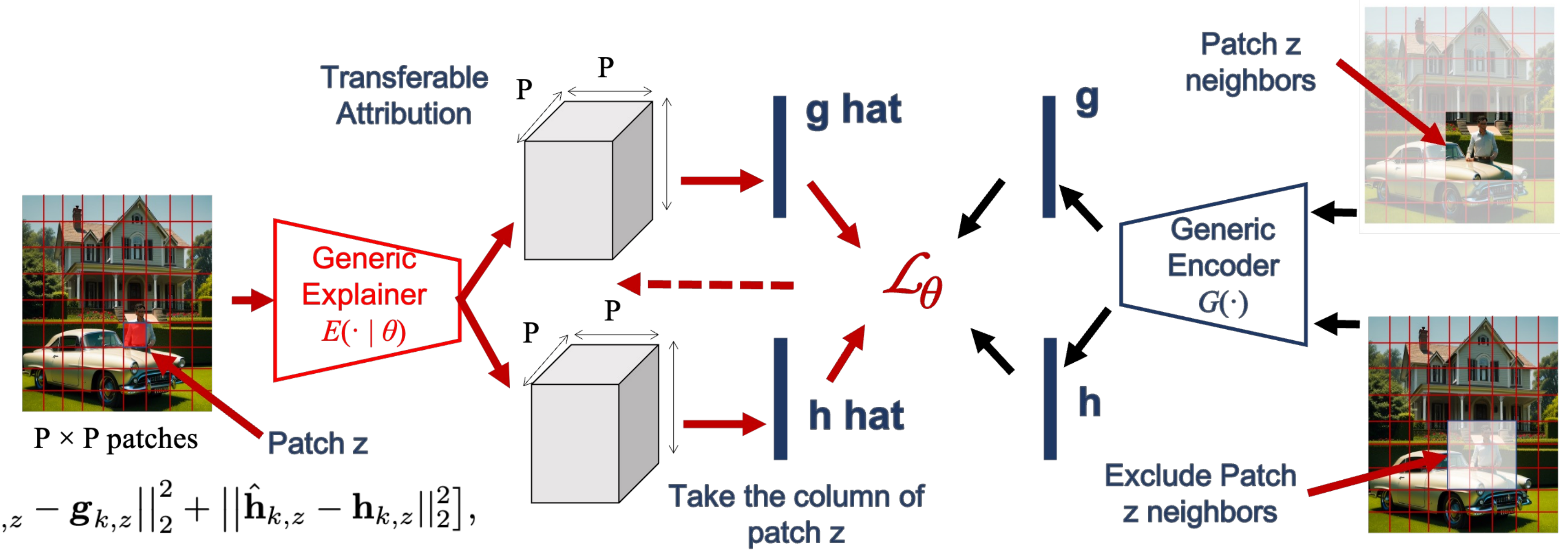
1.2 Feature Attribution can Transfer

The meta-attribution is defined as a tensor that versatilely encodes the reusable attribution knowledge for explaining downstream tasks. In the following Figure, meta-attribution is a three-dimensional tensor. It attributes the importance of input patches to each element of the embedding vector for the meta-attribution. Each PxP slice of this tensor corresponds to PxP patches within the input image, encoding their importance to a specific dimension of the embedding vector. In this way, the meta-attribution inherits the adaptability of the embedding vector, making it versatile enough to adapt various explanation tasks in downstream scenarios.



1.3 Training TVE

TVE generates the meta-attribution g hat and h hat; and updates the parameters of E(• | θ) to minimize the loss function L_θ. The pre-training of E(• | θ) is guided by the meta-attribution instead of specific tasks. This empowers the trained E(• | θ) to remain impartial towards specific tasks, providing the flexibility for seamless adaptation across various downstream tasks.

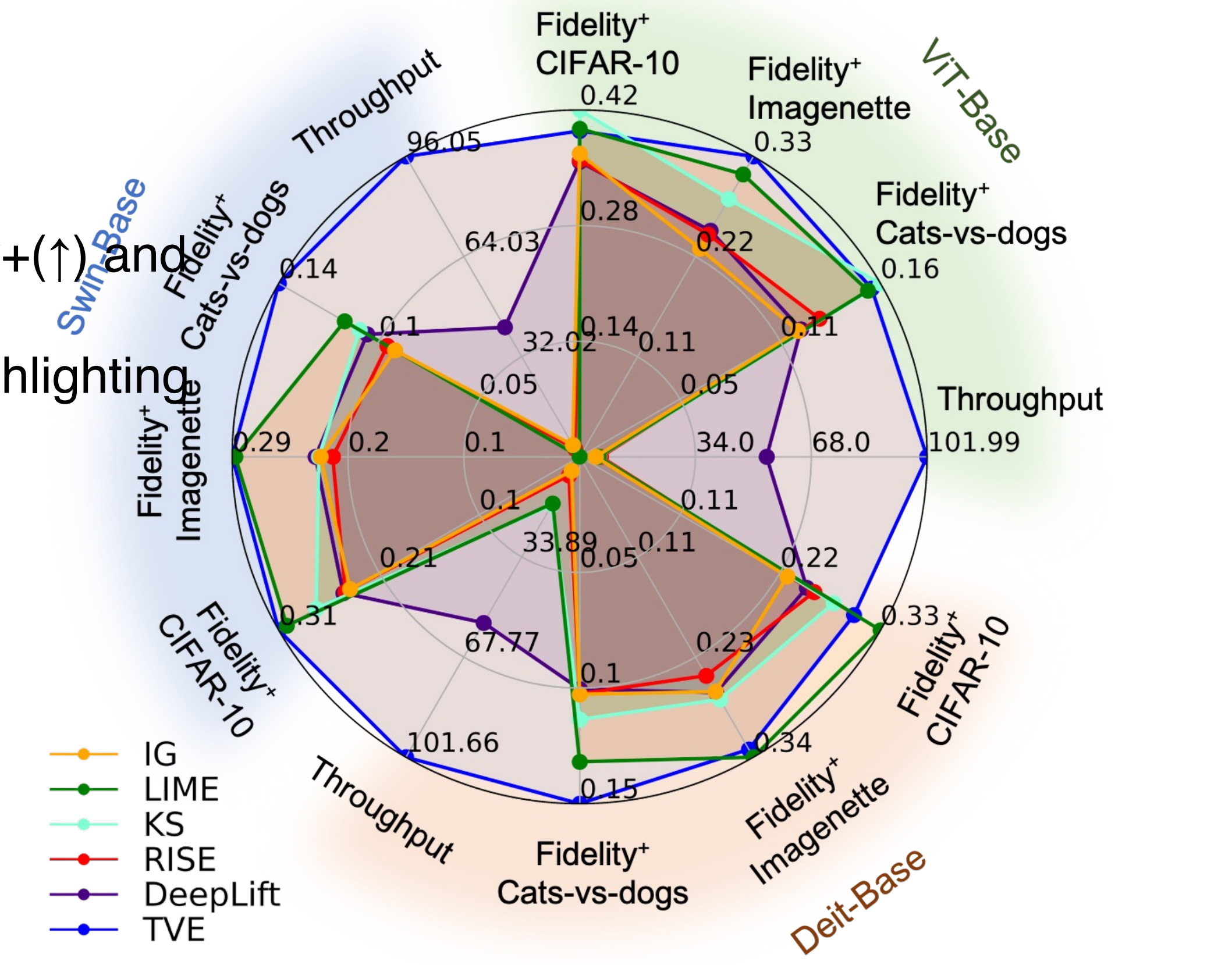


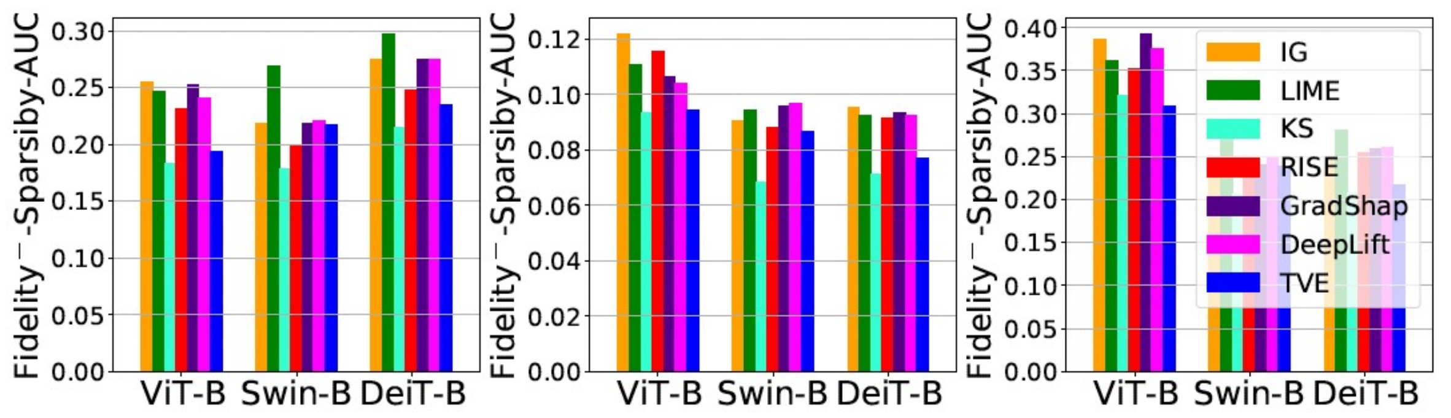
$$\mathcal{L}_\theta(\boldsymbol{x}_k) = \mathbb{E}_{z\sim\mathcal{Z}(\boldsymbol{x}_k)}\big[||\hat{\boldsymbol{g}}_{k,z} - \boldsymbol{g}_{k,z}||_2^2 + ||\hat{\mathbf{h}}_{k,z} - \mathbf{h}_{k,z}||_2^2\big],$$

2. Experiment Results

2.1 Fidelity

- TVE consistently exhibits promising performance in terms of both Fidelity+(↑) and Fidelity–(↓), outperforming the majority of baseline methods.
- TVE exhibits significant strengths in both Fidelity+(↑) and Fidelity–(↓), highlighting its effectiveness in identifying both important and non-important features.





(d) Imagenette (e) Cats-vs-dogs (f) CIFAR-10

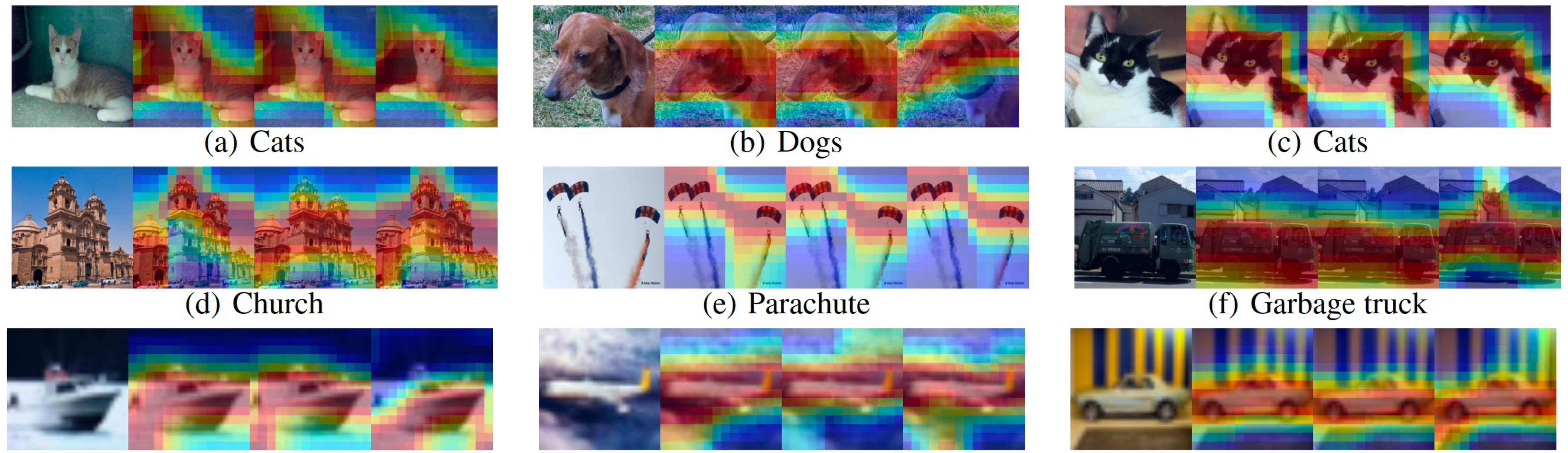
2.2 Transferability

- Both TVE and ViT-Shapley are pre-trained on the large-scale ImageNet dataset, and transferred to the downstream datasets without additional training. TVE has stronger transferability than ViT-Shapley.
- The pre-training of TVE significantly contributes to explaining downstream tasks.
- It is more faithful to explain downstream tasks based on the task-specific classifiers.

	Datasets	Cats-vs-dogs		Imagenette		CIFAR-10	
Target model	Method	Fidelity$^+$(↑)	Fidelity$^-$(↓)	Fidelity$^+$(↑)	Fidelity$^-$(↓)	Fidelity$^+$(↑)	Fidelity$^-$(↓)
ViT-Base	ViTShapley	0.11±0.09	0.13±0.10	0.25±0.13	0.25±0.14	0.36±0.17	0.36±0.17
	TVE-H_g	0.14±0.11	0.10±0.08	0.29±0.14	**0.18**±0.10	0.39±0.18	0.34±0.17
	TVE	**0.16**±0.13	**0.09**±0.07	**0.33**±0.16	0.19±0.12	**0.40**±0.18	**0.31**±0.16
Swin-Base	ViTShapley	0.09±0.05	0.11±0.07	0.24±0.07	0.24±0.09	0.25±0.11	0.28±0.14
	TVE-H_g	**0.14**±0.09	0.10±0.07	**0.29**±0.08	0.24±0.07	0.26±0.12	0.27±0.13
	TVE	**0.14**±0.10	**0.09**±0.05	**0.29**±0.10	**0.22**±0.06	**0.31**±0.14	**0.24**±0.12
DeiT-Base	ViTShapley	0.12±0.08	0.1±0.07	0.22±0.09	0.29±0.11	0.28±0.13	0.24±0.13
	TVE-H_g	0.13±0.08	0.09±0.06	**0.33**±0.10	0.25±0.08	**0.32**±0.14	0.24±0.13
	TVE	**0.15**±0.10	**0.08**±0.06	**0.33**±0.10	**0.24**±0.08	0.30±0.13	**0.22**±0.12

2.3 Heatmaps

- The salient patches emphasized by TVE's explanation reveal semantically meaningful patterns.
- TVE does not rely on pre-processing of the image or post-processing of the explanation heatmap.
- Different model architectures make predictions based on distinct image elements.



(a) Cats (b) Dogs (c) Cats
(d) Church (e) Parachute (f) Garbage truck
(g) Ship (h) Airplane (i) Automobile