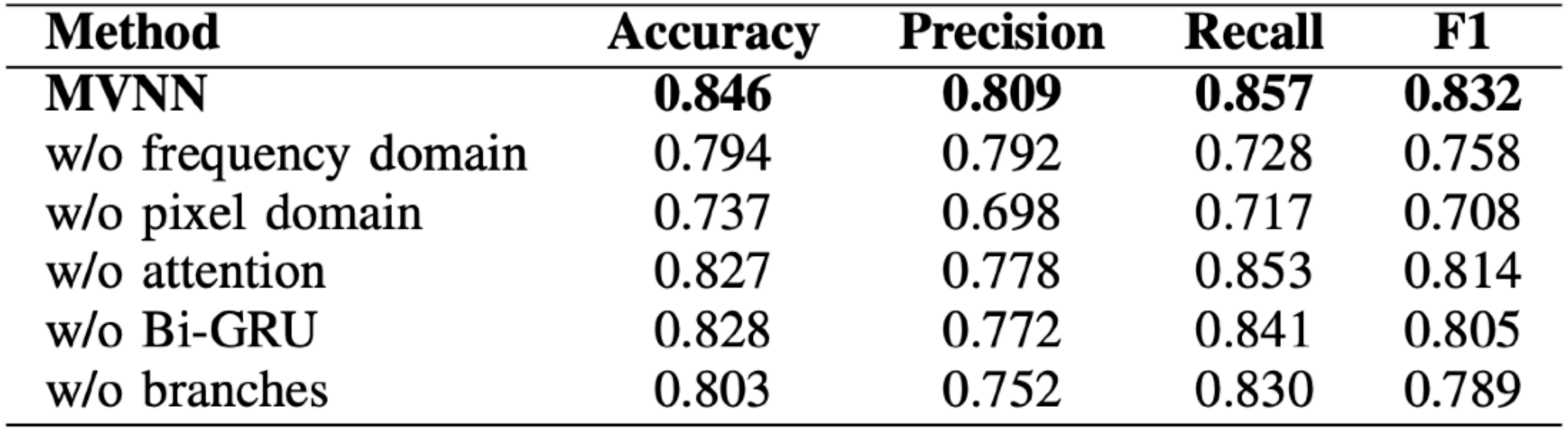
Experiments.......

Ablation Study Observations



Method	Accuracy	Precision	Recall	F1
MVNN	**0.846**	**0.809**	**0.857**	**0.832**
w/o frequency domain	0.794	0.792	0.728	0.758
w/o pixel domain	0.737	0.698	0.717	0.708
w/o attention	0.827	0.778	0.853	0.814
w/o Bi-GRU	0.828	0.772	0.841	0.805
w/o branches	0.803	0.752	0.830	0.789

- **Multiple domains:** The frequency and pixel domain both are important, the accuracy drops by 5.2% and 10.9% without the frequency and pixel domain sub-network. Pixel domain plays a major role and the frequency domain is auxiliary.

- **Network Components:**

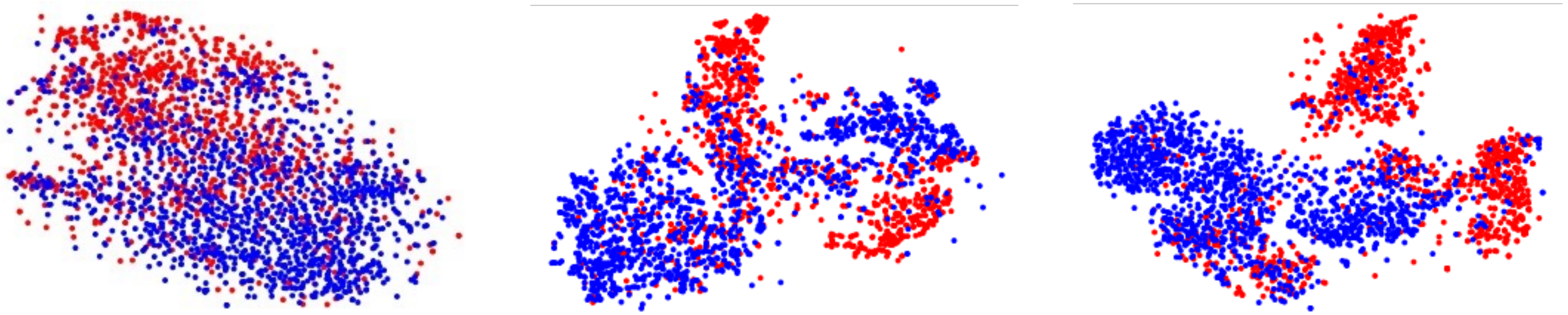
  - remove attention the accuracy is drops by 1.9%, which means that the attention mechanism better than simply concatenating

  - remove the Bi-GRU reduces 1.8%; remove the branches drops by 4.3%.

- Incorporating different levels of features and considering the dependencies between these features both help capture the semantic characteristics of visual contents

Experiments........

visualize the visual features



(a) Frequency domain sub-network (b) Pixel domain sub-network (c) MVNN

- t-SNE show separability of the feature representations: MVNN > pixel > frequency
    - **frequency domain:** positive and negative feature samples overlap a lot
    - **pixel domain:** can learn discriminable features, but the learned features are still twisted together
    - **MVNN:** there is a relatively visible boundary between samples with different labels
- Pixel domain is more effective than frequency domain in distinguishing
- **Fuses information of multiple domains can more distinctive** feature representations, better than single domain