

ABSTRACT

Methods, systems, and computer program products are provided for shared latent space-based debiasing. An example system includes at least one processor configured to: transform data from each of a target domain, which lacks protected features, and a separate source domain, which contains these features, into correlated latent representations; jointly train a cross-domain protected group estimator on the representations; and debias a downstream machine learning model an adversarial learning technique that leverages the group estimator.

<FIG. 4A>