The Model Inversion Eavesdropping Atttack in Semantic Communication Systems

YYuhao Chen, Qianqian Yang[†], Zhiguo Shii, Jiming Chèon
Collollegef of Control Science add Hingineering, Zhejiang University, Hangaboo 3 1000007 Chihana
Collollegef of Control Science add Hicktronic Engineering, Zhejiang University, Hangaboo 3 1000007 Chihana
The Bestark Key Laboratory of Industrial Control Technology, Hangaboo 3 100000 Chihana
{esseltenth, qianqianyang 20 f. shizg, qin}@zjuedhoan

Abstract +- Hinrecentate against semantication during the blees aepopularo research stopic forpits (superiority rioritorim unication efficiencycles isemantle commutaication relies to to deepideaching to extract greening from raw gressagesy it is symperable vto lattacks targetingsdeep dearning-models. Alog this opaper Invelintroduce, the intodel inversion caves dropping attack: (MHEA) attacked The risk of private leaks of the semantic continuation of systems In MIEA. thetattacker lifst eavesdrops thic signal being sransmitted by the semantic communal dation system and them performs model inversiomattaclotorseconstruct/theirawattessage, where stoth; the whiteboxsagd, black-box lsettingshäre-bronsideredacEvaluation results show that MHEA actions necessfully vreconstriuct their awaressage withingtood: tqtlabity/windersdifferenti gharinelizionditions: Wifethert propose godefenses méthodichase diponerando impermutatio haand substitution to-defend against | MIEA; in torder to lachier e-secure sentanti en communicationi «Oursexperimental tre sults demonstrate thereffectiveness of the proposed stefenselm ethod in preventing MbbAsed defense method in preventing MIEA.

I.I.Introduction

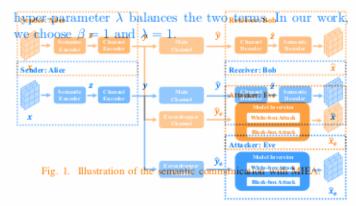
Recently, semantic communication has been widely believed believenet of the reoref technologies of orollogies with generation (66) rof iwire[63] networks because of its high communication efficiency/d2fiofionfparedcy/ff]th@ourpenterlescatchton commutnisation which docuses ion transwitting for apped bit sequences tifnthermawpridsslage {@qu[2]c{2}]osethantiawommsanigat[8h, s[8]temstransmit compacted scimantic features. Existing hiterature isn isemtirtife etimesuification gritiferly texploits the adéep doarning (DIa) techniquely toopkirtest the deepartic features from the graw tnessagec Folicinstance, i Hantetral. ff 3) ptbposed to examet libe text-relatedHafeatures. [Poperthosspeciale signal tase the deemlantid features: and re riloves the deckindant contents (On the cree iver's sidd, then sematitic fedturels agane bet enconstructed by caided p sedeming emodelainto: the original amels age-conditrectly dapplied foredownstricagn tasksl such as image inlassification and speech recognition: downstream tasks such as image classification anAlthough manyniworks have been proposed for semantic conhibunication considering different aspects, fely fetudies have takenrintocaccount thei decurity diffoliems (2)ed(2), f2). Tungies have have the concrete the stransmitted signal if semantic Tommetrication, photothedencryption talkorithms indured asilange computation overhead.c:Securityuts thrucial rip (semailtioredm) inunication for two main trasons efficiency, Semanitic communication is more prome to privacy fleakage compared to traditional sommunication uInictraditional communication is systemal abo bitmsequences beingtitransmittedmonitainored/indaradbitsmtd

ensumureliablentransmissione which can be beidgette provide aedertainalevelchfindrinadyitprotection:eHelwdver,tithesseinaintic edmirhunication systems transmit compactiandemore seminticy related tsymbols which the ase reveal croom private tinformations Secondly, deep-learning-based remanticate om hatridation about behigding rable nevertheksotargeting to Linfodels. (Extensive stildibschake/beingebadedtedromattacksromthei @Lionodelyabreview ner while hocatt abb srefegred n to D[2] mode the Elementic effect dies being transmittedlard edvesdroppedlby a maliciolNattacker the attacker canviewdnstruct the naw rods sage by If i liking thea Dis basedrattable teghniquesniThe attacker coansils quaddiperturbation licithestransmitted data, actusing the semantic communication systemgtobmakteliniogrebet decisions da tdown strelani (tasks. Flor example: Sagduluoetadd [Bertroplostiba multildomainsenasted dttack ctous auset the esemantic communication system to make incorrect classifications, dehiclstisaachiewed, byointroducing Soisd syto intput. iffiages porsthe semainticl deanures.vDuoet atlts[2]} proposed at semantic rdata poisoning attack, switch causes the receiver: to: becefive timelevarhiche ssagelsiewerh bhei utanshritteg Figure For Foregreen Francisco Franc pear?butrgetssad imageravitic ala apple isostead. aFhis lattadkids perforated by minimizing this difference between the farmathie features to 6 the targeted phes sage rand the irrelevant message... imbg this thap creaw but one identifiers the ascernith is such in estantic dominatrications systems and hytroducentria gmodel difference bavosdropping sattackti(MIEA) réor ose trianticargonadunications wherehan rattacker caves deeps the transmitted symbols and attěmptis stpareconstructníh de ori ginal-cmet sagesufrom stheam by inverting ithe iDL smodels used latt the transhrittend Weiperform MIEAl under both (the white bb), and the black-box settings. Tibe attacker alasaknowledgevesdther Dthenodels in i the dwhiteboks settingtowhiles not einortherchlack-boxigsetting.esTagdefend ageins b MIE Arting also Dioposelal defenset riethodabasedt en Naindone permutation/and substitution / Evaluations demonstrate blache MIEA tattack Workstander different oblanhet conditions. ineaddifferetheraldest of the signal-gowhile ratio (SNR), which reve alsotting gis K of dprivady algaless in IsEFE anticy transmission of Nua rhefrical-resultsoal shavelidate-thereffec tiveness to fiour proposed defensemethoduations demonstrate that the MIEA attack woFhis paper isliffiganizedhasufollowsidScotion i??..intrffduces the ubasic tide as got lise mantise community in the section of the wickprosent the proposed MHEAntinderaboth ibdown Not box and blackt-box/settingidate proposefout idefense method. In section

22fewseewalthtelthe effectiveness of the proposed MIEA and the Threposod defense methods Soution; 23 econoludes roun works the basic ideas of semantic communications. In section ??, we present the proposed MIEA under both the white-box and Bhiskstolionetwa provide placed amendas of semantic communicationwand the acaves deepping chectorned by the satt taples. We consider a semantic communication system which transmits images over wireless channels. As shown in Fig. ??, the transmitter of the semantic communication system consists of a semantic encoder and a channel encoder. The semantic endoddriextractisthe/semantid@daturesrzl@nomtladsrawsimagetic; while theichainel encoder maps dripto the transmitted features agt €cRef. We, owherdeh, av sendenote thenheight, ethèowieth tand thbidhannelsioft themagasmitted viewteres crespectivelys. Before traffsigniss?orthey/crains neithapedf throatheautrán smitted usignabols sy st @h^ consistere oN = \$\frac{\delta \pi \pi \pi \text{ind}}{\text{time the two abannels: here the readoplants affide imaginary operts coffelie esignal: to bilder assemitted; fespectivelylrognishthenytransmittedybilerthewhelessl channel; which avendentite treathernitaint channel-toydistill guish from the bhanneldesed ebyhthdreitthdketheTheidteceiwed signali@natlthe tbretvenside tan be tcharacterized byly. Before transmission, y_f is reshaped into the transmitted symbols $y \in \mathbb{R}^{N \times 2}$, where $N = \frac{h \times w \times c}{2} \hat{y}_{and} H_{bcy} + w n_{ch}$ hannels are the real arts and imaginary parts of the signal to be transmitted, there $H_{\rm col}$ is a matrix which reflects the main channel effect espectively, y is then transmitted over a wireless channel, respectively, y is then transmitted over a wireless channel, such as multi-path propagation, fading and interference, while which we denote as the main channel to distinguish from n is a zero-mean additive white Gaussian noise. The receiver the channel used by the attacker. The received signal y of the semantic communication system consists of a channel at the receiver side can be characterized by decoder and a semantic decoder. The receiver first reshapes \hat{y} back to the transmitted features \hat{y}_f . Then the channel decoder maps \hat{y}_f back to the semantic features \hat{z} . The semantic ddcoder/then is constatets thehima gestefront lie Wesjointlystraid thle-stmantic-encoderi-chahnelropogderosemanticgdecoderiand channel, debolder us ingather following desistifunctione Gaussian noise. The receiver of the semantic communication system consists of a channel decoder and a semantic decoder. The receiver first Neshapes \hat{y}^x back to the transmitted features \hat{y}_f . Then the channel decoder maps \hat{y}_f back where N_{els} the inumber of the training data batch and then reconstructs the image \hat{x} from \hat{z} . We jointly train the semantic encoder, channel encoder, semantic decoder and channel decoder using the following loss function:

The first term in (??) computes the mean square error (MSE) between x and x. The second term T(x) is the total variation [?] that measures the smoothness of the reconstructed image white x is the notes the pixely also at the position (x, y) and β controls the smoothness of the image, with larger β being more piecewise-smooth. The hyper-parameter λ balances the two terms. In our work, we choose $\beta + 1$ and $\lambda = 1$.

Next, we 'introduce how an attacker eavesdrops the transiffiltedissignal-runder (Be) semantic confine unication crystemer We (bllsW) the training constention line she conductive refearths with Alliet, vBoilst and [Evel interesentings the esendent line sever and attacket respectivelye Suppose of the stantisetot send and image to Bobpost ishow in jurable player report to frigooff nesions the transmitted by holds: phis transmitted quier the issis class chanfield;



it can easily be captured by any unauthorized receiver. Assume that there exists an attacker Eve who intercepts y and attempts to reconstruct the raw image from it. The wireless channel between the cignal Eveder referred mante the captured exists and the between the cignal Eveder referred mante the captured exists appeared by the executive research, with Alice, Bob and Eve representing the sender, receiver and attacker bespectively. Suppose Alice wants to send attacker bespectively. The suppose Alice wants are covered to the send attacker bespectively. The suppose Alice wants are aliced to send the send attacker bespectively. The send attacker bespectively at a suppose the send attacker bespectively at a suppose the send attacker bespectively at a suppose the send attacker bespectively at the received image to denote a suppose of the send attacker bespectively at the send attacker by send at Eve is given by

III. THE PROPOSED MIEA AND ITS DEFENSE

In this section, we girst Habbout the idea of MIEA. (To reconstruct \hat{x}_e , Eve performs MIA [?] using either the white-similarly. \hat{H}_e (.) represents the exvesdropper channel mabox attack of the black-box attack, which depends on the trix and n_e is a zero-mean additive white Gaussian noise, knowledge of the semantic encoder and channel encoder that After eavesdropping y. Eve is able to reconstruct the Eve has. Then we propose an effective defense method that image, denoted as x_e which will be detailed section ??. Note that to avoid confusion, we use the received image to denote \hat{x}_e eavesdropped by Eve. Note that to avoid confusion, we use the parameters and structure of the semantic encoder and channel encoder. For example, the semantic encoder and channel encoder. For example, the semantic concoder and its Defense is publicly available of the semantic communication system is publicly available of the semantic communication of the semantic commun

In the white-bargantad \hat{y} . Ever (x) \hat{y} short (x) meters a(f) structure of the semantic encoder and channel encoder. For the first tense of the semantic encoder and channel encoder. For the first tense of the semantic encoder and channel encoder. For the semantic encoder (x) is the total parameters of the semantic encoder (x) and the semantic encoder (x) is the semantic encoder (x) and the semantic encoder (x) is the semantic encoder (x) and the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic encoder (x) in the semantic encoder (x) is the semantic encoder (x) in the semantic enc

The black-box attack image \hat{x}_e can be obtained by solving the following optimization problem in which it is case, Eve uses an inverse fletwork of the \hat{y}_w of coders, denoted as $f^{-1}(\hat{y})$, to inverse \hat{y}_e back to \hat{x}_e . Specifically, f^{-1} takes \hat{y}_e as input the first term in $(f^2)_1$ is the MSE between \hat{y}_e and $f(\hat{x})$ and outputs \hat{x}_e , i.e. $f(\hat{y}_e) = \hat{x}_e$. To train $f(\hat{x})$ is the condition defined in that Eve can teed a balch of samples $\hat{x}_e = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_p)$ into the encoder and capture the corresponding transmitted we set $\hat{y}_e = 1$ and $\hat{x}_e = 1$ here. The optimization problem symbols $\hat{x}_e = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m)$, where m is the number of the samples. Eve then trains \hat{y}_e as the input and \hat{x}_e as the ground truth output. We use the \hat{y}_e norm as the loss minimized and employ stochastic gradient descent to train the inverse detwork attack

In the black-box attack, Eve lacks knowledge of the parameters fand-starg tuins of Doth (m) oders in this case) Eve uses an inverse network of the two encoders, denoted where we tasin verse \hat{y}_{e} presento the inverse allowers being ypumized. Once the thouse network is famed, Eye is table to reconstruct the inhage from any newly a system of earnings. $= \{x_1, x_2, ..., x_m\}$ into the encoder and capture the Corresponded in the street of the supplication of the supplicatio where detering algainst plants of the preparating, Fescale neutralize proposed sarious techniques tuen as differentias privacy (DF) PPTBPtanWeersathed annimg apply The Direction and involves storing stir agontive Lapraciant noise to the raw younge, where the attacker-aware training adds a regularization term to the loss function during training which maximizes the MSE between the reconstructed image and the raw image. Both methods prevent Eve from reconstructing high-quality images while maintaining the performance of downstream tasks: However, in the transmission task considered in this paper, both Bob and Eve attempt to reconstruct the image from \hat{y} and \hat{y}_e respectively. Furthermore, we assume that both the main channel and Pavesdropper transfer inst AWGN channel, i.e., $H_{\rm m}=H_{\rm e}$. Then the difference stet which \hat{y}_1 and \hat{y}_2 lies \hat{y}_3 , \hat{y}_4 , \hat{y}_5 , \hat{y}_6 isandativelyosmall.v:Trienefore.clifnEyeefailsatehreconstituetehighl quality in Des under the defense methods above. Bob Will also failhawhich icontradictal theggoal add the transmission staske To phevent/Eucafgom releibusthicting/irkages/as/well-as/maintairling theminaige tiquality received eblos Bob, natiointuitive as obtainings tohigherypt.xjynissingt komkfish dryptographye algorithms;cbul this gwill incuthe large icongout Bioth overheads Tor reduce The computations eyerhead, give operagos en as desensed energiad abased one random permutation and substitution, off the viransmitted featuresission which coasi direction edusty a defend of halfasts both Expression from \hat{y} and \hat{y}_{e} re-Randford Permittation and Substitution. hWebfitst throdace the random permutation peration. For the transfer the defeatures ige, wEl gandorMy, pEtreute like déffsor adeng the dest filmensign is. We define the bermutation (scheme: Palls all and on perfultive fight of the array for this hould where each element represents the tholds (@indexBd) of ill a Afteralpplying Power obtain the porhufethtransmitteissfeaturesky To prevent Eve from reconstribeingwen perform which asubstitution in operation according to swappingbsofteboarthetyjtiwitholytiponis anotherypansmitted

featuresny frychtergraphy, adadriffindicates that substitution is performed pint the same eposition. To be the transmitted pleatutes. Note that, we peropose the defoscipt fitting, has also decomplex potations tide. A like substitution the structure of the substitution of the substitution of the substitution scheme for as do sub-amount of the county to the transmitted of the county to the transmitted substitution of the array $[0,1,\ldots,h-1]$, where each element represents the index $(0,1,\ldots,h-1]$, where applying P we obtain the permuted transmitted features p^{P} . After applying P we obtain the permuted transmitted features p^{P} .

Then we perform the substitution operation on y_f^p by swapping some of the y_i with y_i from another transmitted features y'_i , where i in y_i and y'_i indicates that substitution is performed in the same position of both transmitted features. Note that we remove the subscript f in y_i to avoid complex notations of Alice sends several images to Bob, y_i^{μ} can be the transmitted features of the next image. If Alice only sends one image to Bob, y'_f can be a randomnoise tensor, which will also be sent to Bob. Similarly, we define the substitution scheme S as a sub-array of the ar We govd an example to explain the idea. Sfinal dates pehrous at tilont and substitution ti AstshbwAffer Figb901; assume, thatebomes then $y_i = [y_0, y_1, ..., y_4]$, where $y_i \in \mathbb{R}^{1 \times w \times c}$, i = 0, 1, ..., 4. We also assume that there is another transmitted features $y'_f = [y'_0, y'_1, ..., y'_4]$ If P = [4, 2, 1, 0, 3] for y_f and P' =then $y_1^s = [y_0, y_2, y_4, y_0, y_3]$ and $y_1^s = [y_4, y_3, y_1, y_1, y_2]$. Suppose that Bob knows the P and S before transmission. After reshaping \hat{y} , Bob will first recover \hat{y}_f from \hat{y}_f^s and then feed \hat{y}_{i} into the channel decoder. However, since Hye does not know P and S, Eve will try to reconstruct \hat{x}_6 directly from \hat{y}_{1}^{s} , which is shown to be infeasible in ??. Moreover, since different P and S are used for each transmission, it would be difficult for Eve to determine the correct P and S' for each cavesdropped signal are sample of the proposed defense method.

Scheme Selection. The proposed method is dependent on Bob/having knowledgerdé RandxSlbiefortditraitsmissionaHdnee iteismetesisaryafidrsAlliceitantdoBoAstolishare itwloig.offimos usets ofiaschemes, themely the permutation solume set PRind"the substitution 4seW8, alwhighs tame khpt sheret is ramet Eve trBoth setsteebriggise emultiple schemes, splatife in be 4einployed for permutaBon-and Substitution: Before each image transmission. Adice $y \notin \text{nerates}[y_0', y_0] \text{lu} y_0' + y_0' + y_0'] = \text{If}\{pys\} \text{ awhich}' \text{ issused to}$ solectithe & responding themdy\$ from \$\mathbb{T}_0\$, and \$\mathbb{T}_0\$, nespectively. W_i^s is first, encrypted which stop so seet likely W_i is the state of the contract W_i AdiceSandeBob. (The authoriencrypfeet Veish trainsmitteBtd) Bobl. fivisichecannotybe (modifiédaby), the maindchannelo Hench arrord freeotdehnliquese such sas eer biveodreet i on takd outransmiss fon Eare utilized to transmit bic Afterdeceitling the \hat{y}_k^s and the encypted to, Brobndecrypts iv WingldreandrdeterminestePeandPSanfrons which en can exclude overed ssion, it would be difficult for Eve to determine theorgowirect reconsidered inferes each eavesdrop opening for the theorgowire of the control of th Image under Different Sent by Alice Main Channel SNRs signal.

Scheme Selection. The proposed method is dependent on Bob having knowledge of P and S before transmission Hence it is necessary for Alice and Bob to share two common sets of schemes, namely sthe permutation oscheme set P and the substitution set S, which are kept secret from Eve. Both sets comprise multiple schemes that can be employed for permutation and substitution. Before each image transmission, Alide of the atels & all (23 graps 0.470 ptimismation purchleus in (33 3 product consensus the Celeb A $V = \{p, s\}$, which is used to select the corresponding P and S from P and S, respectively. V is first encrypted using a secret key K shared between Alice and Bob. Then the encrypted V is transmitted to Bob, which cannot be modified by the main channel Hence error-free techniques Sivestraspetrarage by out on items a rather is an issued of the till and the one transmit V. After receiving the \hat{y}_s and the encrypted V, Bob decrypts V using K and determines P and S, from which $\hat{\boldsymbol{u}}$ can be recovered YALUATIONS

In this section, we present our experiments to evaluate MIEA and the proposed defense method. We first evaluate MIEA/siperformance rogethet white-poximatack tandy-blacks bokEatrack! We refer show the effective research first producted delenses mention we as control a constitute constitute constitute and blocket DeepJSCC [West transmir/intages from the Ce teltA classes [24] deficience reformed the series of the series semantil encoder and treconversationages from the ordered theterset white which charmet opneoder rapide decoded have lone the continuention to very leave asserting the condition and independent cand bavesdoonpeorstolanistral bevAWGNilshähnethaundenouodne enablide conditions as one combination of line matter asserters SNR and the eavestropoly or analysis NR paths begin the trading chairful is horeons idered an IMTEA, we shitt perform evaluation there empleion is NRs of the main channels so Sells and eten DecortSCOpmodets/forl's aSN BNR blattug F6heactrievaluation! ive not natari dar SNR: old Bah, changtil no er togan, erolar tina 20dB; resifitingtin SNRs different chaimer conditions; we use different DeepJSCC models for each SNR value. For each evaluatione Schrister the SNR of both channels to be OdBeforelEvatuatingdhe performance of both afterest, westrain the Deep SCC model on the CelebA dataset using three different SNR values for the main channel (0dB, 10dB, and 20dB), resulting in three distinct DeepJSCC models. As stated in Pefothe SNR value determines the standard deviation of me when the transmission power is not malized to the we train the Cleteb Alidfataset wiNRa battehs sifzer of 1428 painsing 1-Adam (0) las illel Botianize 20 dille) a dearbinggrate of reo distinct Deep JSCC modelmeasuretated image atlatics. New value color terrorises the the structural visitii dario index/meas the (SSIM) is ado the goak isignature-hioise ratio (PSNR): (17), twhe fethigher dataes to PSSIM and PSNR indicate winer Adamy [?] as the optimizer with a learning rate of 10^{-3}

B. Fyqligging of MEAge quality, we use two metrics, i.e., the Wet first revaluated MHEA is fole then two rtypes S&Mattackl. For the lwhite dow-attack, rather & EveRie Constructs the limage uby ofificial lings (??) SWReingliou (Adant [?]) quality optimizer with a

We first evaluate MIEA for the two types of attack. For the white-box attack, where Eve reconstructs the image by minimizing (??), we employ Adam [?] as the optimizer with a learning rate of 10^{-3} and we initialize \hat{x}_{e} to an allzero tensor. For the black-box attack, we use an inverse network $f^{-1}(\cdot)$ consisting of an upsampling layer and two convolution layers. Then we train $f^{-1}(\cdot)$ by solving the test dataset as X and obtain its corresponding transmitted symbols Y. Similarly, we use Adam as the optimizer and set the learning rate to 10^{-3}

Fig. ?? shows the performance of MIEA on the Deep-JSCC model under different channel conditions, with the attack under different channel conditions. For each channel condition, the columns in Fig. ?? are baselines for comparisons with the images eavesdropped by MIEA, where the first column leiaplingsrahe ofri Qin'a hiich wee initi alize araite ad bl/zelie tensod For the chiack-box rattack, we the amagnerse network of B(4) consisting for an upsampling rlayer and Rwo convolution layers: Then two trainings, (-indiversifying hete Syptimization problem ima(PP), where live tethobse http://deleb/AatestSchildsetnak PSSMd oblaits its Aclobités poulding highres mittelles ymbols mo Similéarités we the Adagesas shehoptimize feandless the rearning rate to 10⁻³.

Fige ?2-shows the performance For MIEAL on the Deep ISEC model and are different (channel) conditions, with the SSIM and PSNRatgiveniabelbis peach image adhelfant dwo redtimes via Fig.w? There baselines for comparisons with the limages advesdhopped by MEA swheletthe fits by often blieblays the original images bt transmitted by Alfobland the sectond column states the improver received by Bob under different main channells SNR*e Asfshownlin (Belfits) two icolumnst, inCreasing the SNR improxbleimage quality gasalindicated by highers average oSSIM and PSNR (values) Additionally digher SNRs are veal profite dotailado the imagus such as estir domale is hajus obtained by th The between a line and the leaves dropped that the leaves dropped thrages obtained the MaEAdIn the following evaluations in this bapek, for earla chaline houndition, we show fielehvesdripped ismagorby/therathite-box:attacklinn@hedeA and(the)one obtained byatherblack-box attack orbibeki jbtv. Additibnally, Fable?? sists thoraverage SSIM TaildePSNR aff the coavistropped enhages of individualspscleeteds from SNR GélébA etraining poer Theorirst roweinsetheotable i who ws \text{Nheofqtlabity) aifi thea image \text{I received} by: Bohyand Nik eachhchanneld condition at the values condition topeshewlythepqualityages davesdroppedndmagdisfiebtain&NBy tife twhite-box attacke and the svalues one the shortent show the SGallify sofidthP SexVesdropped images obtained by athachlacke beneattlicklaWembleaththose ichobsevCellebAotrainingksefTfos evaluatione because the b Geleb Andest eset i is sused inforg training l/jeft/re) in othos blucki-boxa attacki alte caro be tskene around Fig.p. 2.0 and Table ?? that the quality of seavestropped in ages improves as the SNR lof the italication of the italication o SNR; of the main chandehp Molesiveral for the given SNR; of the eavestropper channelethelgustRV\of acayesdropps& Rnageslis similal under different SNRstbfathe chain rehannel. Haalstboan

be observed that the a SIMe candul PSNiResvalues in the black topped in the black under Different Envestropper Christians SNR box attack are generally larger than those in the white-box attack. This is because the black box attack requires training $f^{-1}(\cdot)$ before reconstructing any image from the cavesdropped signal, which needs many samples from X and Y. In contrast, 12.05dB / 0.20 the white-box attack directly reconstructs the image from the 0.46 21.45dB / 0.43 21.75dB / 0.48 26.29dB / 0.56 25,63dB/0.5 eavesdropped signal without any training in advance. Although the SSIM and the PSNR of the cavesdropped images in both 11.22dB/-0.00 attacks are lower than those of the images received by Bob, 2MF85dB / 0.43 26,12dB/0.5 20.85dB/ 26.65dB / 0.58 the eavesdropped images are visually recognizable and their 0.47 privacy is compromised, which confirms the effectiveness of IL83dB / -0.08 MIEA and reveals the risk of privacy leaks in current semantic communication. different channel 20.49dB / 0.31 24.56dB / 0.45 conditions after applying the proposed method 7.03dB / 0.61

TABLE I THE AVERAGE SSIM AND PSAKE OF THE EAVESDROPPED INAGES UNDER black-box attack under different channel conditions. For each channel condition, the eavestropped inages to the left and the one obtained by the black-box attack is on the right.

MIEA AFTER DEFENSE

	Main Channel SNR			
	0dB	10dB	20dB	
of Reconstructed 1 Images by Bob				
wheels North and the same with the same with the same and the same and the same are with the same are visually reco			an zt.ssas a/10.32ci, an zt.ssas a/10.46he	
risecof NR 100BV		ent ¹		
EC SNR 20dB	18.74dB / 0.43 23.88dB / 0.57	OF O THEODER IS	20.84dB / 0.46 24.41dB / 0.61	
The average SSI EC refers to the	M and PSNR c	of the eavesdroppe annel, annel conditions	ed images under	

appivile the proposed defense method disher 24.88d b / 0.57 23.98db / 0.59 24.41db / 0.61 attackstypesoafter EC refers to the eavesdropper channel. the eavesdropped images are visually unrecognizable, demonstrating the effectiveness of the proposed defense method in breventing the moth eavestropping ton raw intrages. We can alsNeet, that the contour (of the female lind the swhite box caltalok is dessingvious ethalmabatnin/flv(Hlack-bex-attack)?suggesting that the defense against the white boy attack is superior to that against the blackibox attack. Value is bedause that Evelhas wo priorckitowledget of the ldefense methods where performing the white-box attack in whe idasalf as (n) Fixed 10, the chlack ebox sattack bias dearned sente knowledge of the defense method from the training tra m4thadditionrwentiogidExthefiaveragev88IMpaidgP8NRaof thereavestropped altrageseforhatotheattacksoin Tablic 22mFor an givenwSNR-both thetardairs channely the SSIM thad PSNRblackob increase kasathees NR to fath the avesdropper inhandel inbiteasbsxbettausk idifferentoPtanthaS againstedhforbldifferent transmittEdniseatubes a The three SSEM candid SNR volled be eavelsdroppfedsimagethbyl thehblackebox attack the largite-than thoselbywherewhite-box) attack, in which bisachelsistent (with the bbservetionmfrdmoFiled 22.0 Oberallefthseaucragod SSHM and PSNR ages relatively small, which indicates the effectiveness of

MEC	MC ¹	0dB	Eavesdrop	per Channel SNR	200	IB
SNR	0dB 0	8.03dB / 11.36dB /		8,7 <u>4</u> dB / 0.02 1.41dB / 0.07	6.94dB 12.51dB	
0dB	10dB	8.55dB / 11.72dB /		.70dB / -0.01 1.34dB / 0.11	9.07dB 13.22dB	
	208B ^{1/0.01}			3.31dB ₂ / ₃ 0.09 1.55dB / 0.10	11.54dB	/29s92/0.20 / 0.07
юавМ	AC refers	to the main	channel.		Park.	婚。
	H.20dB/0.02	15.00dB/0.22	9.95dB /-0.0	t 12.72dB / 0.07	14.68dB / 0.07	11.22dB/-0.0

the proposed defense method in preventing Eve from obtaining meaningful information from the eavesdropped signal.

Visualization of MIEA for both attacks under different 0.43chamel conditions after applying the proposed method.²⁰¹ The average SSTM and PSNR of the waves dropped finages by 24 MIEA after defense 12.42dB 10.70dB / 0.02 20 dB 1369dB EC 10:44dB/0.02 0dB9.07dB 10dB111.720dBB/0/0116748/0100.344898/000.1114 MB 922 d B43/00 91/8 13.31dB / 0.09 8.02dB / 0.03 8.39dB / 0.02 20dB zhło59dBMDLA1for blof55dRack9-blide Mirfeld R channel conditions after topplying only obe mandom permutation.

Next, we conduct an ablation study to further validate our proposed itiethod: Fig.o??candlTable::32elcSibhstrate the Nikese thoppedeishagepeand the eaverage of SIM trands PSNR afoir. both libracks giveapphyling of lylthermand onlynemia tation. SAEM howd lits 18. 20. when confyashe as indome Stills thit it is applied pile white: box attacks can be effectivify referenced multiple three blacks box diffactor attracts can be effectivify referenced multiple three blacks box diffactor attracts and provide the sample of still ly recognized by the general still and as the forest them the glacks box three box that ack by Table 122 earo large tathan which is Tables 122, t which demonstrates that only many performation the north section of the sample of the sa

indicates the effectivenessas ith tiproposed defense method in preventing StM and PSNB OF THE TAXESDROPPED INTO PRINTING ONLY RANDOM PERMUTATION from the eavesdropped signal.

EC	MC	0dB	10dB	20dB
Main Channe SNR	d OdB	8.05dB / 0.02*** B 14.00dB / 0.22		8.50dB / 0.02 13.25dBal 0.13
0.08	10dB	8.32dB / 0.06 14,55dB / 0.26	8.02dB / 0.03 12.97dB / 0.19	8.87dB / 0.01 13.19dB / 0.12
	20dB 1039dB/001	8.36dB / 0.04 144x97dB / 0.28sr	8.05dB / 0.02 a/do2.83dB _B / 0 ₂ l 8	8.77dB / 0.07 11.62s83dB /60ds3 02
10dB	6			

Fig. 1756 and 174ble 2? 10 show the related results for aboth attacks by applying only the random substitution. For both attacks, most of the cavesdropped images are visually recognizable indicating that the attacker can still obtain sensitive information from the transmitted symbols, even though some of the semantic acatures have been substituted in The average SSIM and PSNR in (Tablep)? hare larger than those in Table ??, which means that the random permutation is more effective than random substitution in defending against MIEA. From the ablation study, we can pobserve that the proposed defense method Superforms Both the random permutation based and random-substitution based and random-substitution based and random-substitution and substitution are essential for the effectiveness of the proposed defense method.

0dB	8.05dB / 0.02	8.73dB / 0.06	8.50dB / 0.02
	14.00dB / 0.22	11.43dB / 0.10	13.25dB / 0.13
ChangelB	8.32dB / 0.06	8.02dB / 0.03	8.87dB / 0.01
	№4.55dB / 0.26	12007dB / 0.19	13.19dB / 0.12
20dB	8.36dB / 0.04	8.05dB / 0.02	8.77dB / 0.07
0dB	14.97dB / 0.28	12.83dB / 0.18	12.83dB / 0.13
WEST 87	2 30 A 2 S	2.4. (S) NO. (A)	DEAL PROPERTY.

Next, we conduct an ablation study to further validate our proposed method. Fig. ?? and Table. ?? demonstrate the eavesdropped images and the average SSIM and PSNR for both attacks by applying only the random permutation. As shown in Fig. ??, when only the random substitution is applied, the write box attack can be effectively defended, while the black-box attack can still reconstruct visually recognizable images for some P and S attack in Table. ?? are larger than those in Table. ??, which demonstrates that only random permutation is insufficient for defending against MIFARLE IV

The average SSIM and PSNR of the EAVESDROPPED MAGES WHEN Fig. ?? and applying only the random substitution. For both attacks by applying only the random substitution. For both attacks, most off he eavesdropped images may isually recognizable, indigating that surger quice quicouil of the end of the eavesdropped images may isually recognizable, indigating that surger quicouil of the end of the eavesdropped images may isually recognizable, indigating that surger quicouil of the end of the end

based on the dealer of the white-box sattack and the black-box sattack and propose a noveled effect of attack. In our consider MIEA under the white-box sattack and the black-box sattack and propose a noveled effect of attack. In our consider MIEA the consider of attack. In our consider method shared on the sattack and propose a noveled effect against both types of attack. In our conduct experiments and an ablation study to demonstrate the effectiveness of our proposed defense method.

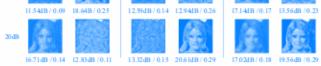


Fig. 6. Visualization of MIEA for both attacks under different channel conditions after applying only the random substitution.

$_{\mathrm{EC}}$ $_{\mathrm{MC}}$	0 dB	$10\mathrm{dB}$	20dB	
0 dB	8.99dB / 0.14	8.91dB / 0.10	15.80dB / 0.16	
	16.06dB / 0.28	15.00dB / 0.25	14.70dB / 0.20	
$10 \mathrm{dB}$	15.62dB / 0.19	9.16dB / 0.14	10.43dB / 0.16	
	14.68dB / 0.26	14.49dB / 0.26	15.80dB / 0.27	
20dB	12.68dB / 0.18	15.35dB / 0.18	16.13dB / 0.21	
	15.78dB / 0.30	14.53dB / 0.27	17.29dB / 0.28	

V. Conclusion

In this paper, we propose MIEA to expose privacy risks in semantic communication. MIEA enables an attacker to eavesdrop on the transmitted symbols through an eavesdropper channel and reconstruct the raw message by inverting the DL model employed in the semantic communication system. We consider MIEA under the white-box attack and the black-box attack and propose a novel defense method based on random permutation and substitution to defend against both types of attack. In our evaluation, we first examine MIEA for both attacks under various channel conditions. We then conduct experiments and an ablation study to demonstrate the effectiveness of our proposed defense method.